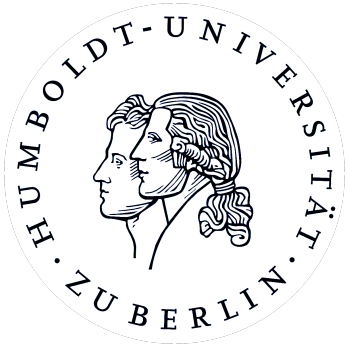
Applications of Advanced Analytics to the Promotion of Freemium Goods



D O C T O R A L T H E S I S

to acquire the academic degree of

doctor rerum politicarum
(Doctor of Economics and Management Science)

submitted to the
School of Business and Economics of
Humboldt-Universität zu Berlin

by

Julian Runge, M. Sc.

President of Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dean of the School of Business and Economics:
Prof. Dr. Daniel Klapper

Reviewers: 1. Prof. Dr. Daniel Klapper
2. Prof. Dr. Stefan Lessmann

Date of Colloquium: September 15 2020

Summary

"Freemium" (free + premium) has become a workhorse pricing model in the digital economy: A basic version of a product or service, e.g., mobile applications ("apps"), can be used for free in perpetuity and premium upgrades are available against payment of a fee. Consumers downloaded apps 194 billion times in 2018 and spent $101 billion on in-app purchases in the same time period. Accounting for almost 80% of that revenue, gaming in particular has seen an unparalleled expansion of demand. It is estimated that 50% of mobile app users play games regularly and that a global total of 2.4 billion people will play mobile games in 2019.

The core thesis of this dissertation is that promotions are essential to the marketing of freemium goods such as mobile apps and games. While freemium already represents a promotional pricing tactic in using a zero price for free sampling, the author conjectures that firms can operate their freemium offerings more profitably by using further promotional tactics, especially targeted and personalized promotions, to sell premium upgrades. The author also argues (and shows) that widespread concerns around the use of promotions, particularly developed in the setting of consumer packaged goods, do not apply in the same way in this setting. This thinking is qualified and developed across four chapters that represent individual papers after providing an introduction to the work in the first chapter.

The work is empirical in nature and applies advanced analytics, in particular field experimentation and machine learning, in collaboration with firms. As representative of the freemium app economy, the collaborating firms observe dense user data that enable the author to both derive insights on consumer behavior that extend existing conceptual thinking in the field of marketing and to devise decision support and expert systems that allow firms to operate more profitably in this setting.

Zusammenfassung

"Freemium" (Free + Premium) hat sich zu einem führenden Preismodell für digitale Güter entwickelt. Dabei kann die Basisversion eines Produkts, z.B. von Handy-Applikationen ("Apps"), unbegrenzt kostenlos genutzt werden und Firmen bieten Premium-Erweiterungen gegen Bezahlung an. Konsumenten haben in 2018 194 Milliarden mal Apps heruntergeladen und 101 Milliarden US-Dollar für In-App-Einkäufe ausgegeben. Beinahe 80% des Umsatzes auf App-Stores wird dabei durch Handyspiele generiert. 2,4 Milliarden Menschen haben in 2019 Handyspiele gespielt, was der Hälfte aller App-Nutzer im gleichen Zeitraum entspricht.

Die Hauptthese dieser Dissertation ist, dass preisreduzierende Sonderangebote von großer Wichtigkeit für das Vermarkten von Freemium-Gütern sind: Obwohl Freemium bereits eine extreme Preis-Reduktion darstellt, indem es ein Produkt Konsumenten kostenlos zum Ausprobieren zur Verfügung stellt, können demnach Firmen durch die Nutzung weiterer Sonderangebotstaktiken höhere Profite generieren. Die Arbeit postuliert weiter (und beweist dies empirisch), dass lange angenommene Risiken in der Nutzung von Sonderangeboten, die vor allem bei klassischen Konsumgütern etabliert wurden, im Freemium-Bereich in dieser Form nicht zutreffen. Diese Perspektive entwickelt und vertieft der Autor über vier individuelle Papiere, die zusammen mit einer einleitenden Zusammenfassung die fünf Kapitel dieser Dissertation ausmachen.

Die vorliegende Arbeit ist empirischer Natur und wendet "Advanced Analytics", insbesondere Feldexperimente und maschinelles Lernen, in Zusammenarbeit mit Firmen an. Als repräsentativer Forschungsgrund dienen dabei Freemium-Handyspiele, in denen Firmen detaillierte Daten über Interaktionen mit Kunden sammeln. Anhand dieser Daten leitet der Autor neue Kenntnisse über Kundenverhalten ab und entwickelt Entscheidungsunterstützungssysteme, die es Firmen ermöglichen, höhere Gewinne beim Verkauf von Freemium-Gütern zu erzielen.

Acknowledgements

First and foremost, I wish to thank my advisor Daniel Klapper for supporting me with steady and rigorous guidance throughout the incredibly educating years of doctoral research. Similarly, I want to thank Michaela Draganska who supported and directed me in approaching empirical enquiry, identifying interesting research questions and creating captivating conference and seminar presentations. Without their comments and feedback, I would often not have entered deeper loops of thinking and analysis that always proved valuable. I also want to express my gratitude to Stefan Lessmann whose research provided guidance and inspiration throughout my doctoral studies and who kindly agreed to act as the second reviewer of my dissertation.

I am further grateful to Jonathan Levav for inviting me for two research visits to Stanford Graduate School of Business. The time I spent there, working in my cubicle next to other doctoral students, meeting with faculty and students, attending seminars and lectures, was incredibly educating and motivating. Working with Jonathan and Harikesh Nair furthered my understanding of the research process and the marketing literature in immeasurable ways. I thank them for the time and effort spent on mentoring me and driving our research forward.

A large part of this work has been possible because I was embedded as a data scientist with companies. At these industry collaborators, I had direct access to company databases and could run field experiments in collaboration with marketers, engineers and product owners. This embedding with industry was invaluable in directly exposing me to managerial thinking and practice and in terms of the experiments I could run and the data I had access to. It was also straining, in that I had to bridge the gap between academic and practitioner goal sets and motivations. Bridging this gap required me to "go the extra mile" more often than not and I am proud to say that many of the analyses I conducted and advanced analytics applications I built led to significant profit improvements for the respective firms. I want to express my gratitude to these companies, especially for allowing me to publish academic papers based on the research I conducted using their tools and institutional knowledge.

Throughout the years of my doctoral studies, I had inflective conversations with people in both academia and industry that helped shape my work and thinking. I want to thank these individuals: Daniel Guhl, Narine Yegoryan, Sebastian Gabel, Vlada Pleshcheva, Lutz Hildebrandt at the Institute of Marketing of Humboldt University Berlin; Wes Hartmann, Susan Athey, James Lattin, Stephan Seiler at Stanford University; Christine Moorman, Richard Staelin and Allison Chaney at Duke University; Boi Faltings and Florent Garcin at the Artificial Intelligence Laboratory of Ecole Polytechnique Fédérale de Lausanne; Oded Netzer at Columbia University; Anja Lambrecht at London Business School; John Langford at Microsoft Research; Dan Friedman and Kristian Lopez Vargas at University of California Santa Cruz;

Rafet Sifa at the University of Bonn and the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS); Anders Drachen, first at Aarhus University and then at the University of York; Ryan Weldzius at Princeton University; William Grosso, Jens Begemann, Jan Miczaika, Neil Young, Eric Seufert, Ching-Hao Hu, Annelie Biernat, Eloïse Pellerey, Sophie Vo, Yavuz Acikalin, Nenad Zivic, Sharon Biggar, Vince Darley, Peng Gao and Christoffer Holmgard at various companies that I had the chance to collaborate with.

I also wish to thank the organizers and participants of academic conferences and seminars that provided helpful feedback for the research underlying this dissertation (in the United States unless otherwise noted): the quantitative marketing work-in-progress seminar at Stanford University; the work-in-progress seminar with my co-doctoral students and the one with faculty at the Institute of Marketing of Humboldt University Berlin; the Theory and Practice of Marketing Conference 2019 at Columbia University; the Marketing Science Conference 2017 at the University of Southern California in Los Angeles; the MIT Conference on Digital Experimentation (CODE) 2016; the inaugural Interactive Marketing Research Conference 2018 in Amsterdam, Netherlands; the Hawaii International Conference on Systems Sciences (HICSS) 2018 in Hawaii; the INFORMS Revenue Management and Pricing Conference 2019 at Stanford University; the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment 2016 in San Mateo; the IEEE Conference on Computational Intelligence and Games 2014 in Dortmund, Germany. I further would like to acknowledge presentations to industry experts that yielded insightful comments: at the Algorithm Hour of Stitch Fix in San Francisco; at the Uber Behavioral Science Symposium in San Francisco; at the Game Analytics and Business Intelligence conference 2015 in London, United Kingdom; at Casual Connect 2015 in Tel Aviv, Israel.

Maybe most importantly, I am grateful to the people close to me who were emotional anchors during this journey driven by curiosity and the desire to uncover new frontiers of data-driven decision making. Finally, I thank the people who proofread or reviewed earlier versions of this thesis and the papers it contains. All remaining errors and the expressed thoughts and opinions are strictly my own.

May 2020,

Julian Runge

Contents

List of Figures

List of Tables

Abbreviations

ANOVA	Analysis of variance
AUC/ROC-AUC	Area under the receiver operating characteristic curve
B2B	Business-to-business
B2C	Business-to-consumer
CA	Canada
CDF	Cumulative distribution function
CI	Confidence interval
CLV	Customer lifetime value
CPG	Consumer packaged goods
CTI	Click-to-impression rate
CTR	Click-through rate
DACH	Germany, Austria, Switzerland
Deep-NN	Deep neural network
DT	Decision tree
EU	European Union
FPR	False positive rate
FR	France
FTUE	First-time user experience
GDP	Gross domestic product
GDPR	General data protection regulation
HMM	Hidden Markov model
IAIS	Institute for Intelligent Analysis and Information Systems
IAP	In-app purchase
i.i.d.	Independent and identically distributed
iOS	Operating system of Apple devices
LASSO	Least absolute shrinkage and selection operator
Logistic	Logistic regression
LR	Linear regression
LTV	Lifetime value
MB	Mega byte
N	Sample size
NN	Neural network
prob(B>A)	Probability that treatment group B's outcome is significantly higher than control group A's; as assessed by a Bayesian significance test
Q-Q plot	Quantile-quantile plot
ReLu	Rectified linear unit
RF	Random forest
RMSE	Root mean squared error

ROC	Receiver operating characteristic
SD	Standard deviation
SMOTE	Synthetic minority oversampling technique
SUTVA	Single unit treatment value assumption
SVM	Support vector machine
tanh	Hyperbolic tangent
TPR	True positive rate
UA	User acquisition
UK	United Kingdom
USA	United States of America
USD	United States Dollar

Chapter 1
Introduction

"Freemium" (free + premium) has become a workhorse pricing model in the digital economy (Gu et al. 2018): A basic version of a product or service, e.g., mobile applications ("apps"), websites or streaming of music and video, can be used for free in perpetuity and premium upgrades are available against payment of a fee. Such premium upgrades are offered either in one-off purchases or in subscriptions. Examples for the latter are The New York Times or many other news websites, health and lifestyle apps such as Calm or Runtastic, streaming services such as Spotify, Soundcloud or Hulu, networking platforms such as LinkedIn, dating services such as Tinder, Bumble or Hinge, or online games such as Fortnite or Roblox (Levitt et al. 2016) to name but a few. One-off purchases are particularly common in online mobile games such as Candy Crush Saga or Clash of Clans. Many mobile games also offer a combination of subscription and one-off purchase options, as does the dating app Tinder that was the leading non-gaming app in the App Store top grossing charts in 2019 (Perez 2019).

Consumers downloaded apps 194 billion times in 2018 and spent $101 billion on in-app purchases in the same time period (App Annie 2018). Accounting for almost 80% of that revenue, gaming in particular has seen an unparalleled expansion of demand, additionally fueled by online social networks that facilitate viral sharing and network effects (Alsén et al. 2016; Sensortower 2019a). It is estimated that 50% of mobile app users play games regularly and that a global total of 2.4 billion people will play mobile games in 2019 (Kaplan 2019). This explosive growth has not only given rise to a large mobile gaming industry that is estimated to drive 60% of overall revenue in the gaming vertical in 2019 (App Annie 2018, p. 20), but to the

prevalence of new types of consumer-firm interactions (Einav et al. 2014; De Haan et al. 2018; Tong et al. 2020).

The core thesis of this dissertation is that promotions are essential to the marketing of freemium goods such as mobile apps and games. While freemium already represents a promotional pricing tactic in using a zero price for free sampling (Bawa and Shoemaker 2004), I conjecture that firms can operate their freemium offerings more profitably by using further promotional tactics, especially targeted and personalized promotions, to sell premium upgrades. I also argue (and will show) that widespread concerns around the use of promotions, particularly developed in the setting of consumer packaged goods (Mela et al. 1997; Jedidi et al. 1999; Anderson and Simester 2004; Günter and Klapper 2007), do not apply in the same way in this setting. I qualify and develop this thinking across four chapters that represent individual papers that are introduced in more detail below and summarized in Table 1.1.

The work is empirical in nature and applies advanced analytics in freemium gaming apps in collaboration with firms. Building on (Bose 2009, p. 1), I define advanced analytics as the tools "used to direct, optimize, and automate [firms'] decision making to successfully achieve their organizational goals." The importance of advanced analytics for industry is closely interwoven with the ascent of "big data," i.e., the availability and storability of vast amounts of data in digital settings (Barton and Court 2012). In the app economy (Arora et al. 2017), firms can observe dense behavioral data describing every interaction of a user with an app in addition to meta data such as mobile device characteristics and geolocation information (Sifa et al. 2018). This pool of data provides the empirical study ground for development of this thesis. Methodologically, I use field experimentation and machine learning to both derive insights on consumer behavior that extend existing conceptual thinking in the field of marketing and to devise decision support and expert systems that allow firms to operate more profitably in this setting (Turban and Watkins 1986).

Table 1.1: Overview of individual papers contained in this dissertation

Chapter	**Chapter 2:** Price Promotions in Freemium Settings	**Chapter 3:** Churn Prediction and Prevention in Freemium Apps: Can Free Premium Goods Change Churners' Mind?	**Chapter 4:** Early Detection of Paying Users in Freemium Apps: An Application of Deep Learning and Synthetic Oversampling	**Chapter 5:** Monetizing Freemium Play: A Practical Evaluation of Pricing Tactics in a Mobile Game
Research objective	Assess the impact of regular price promotions in a freemium setting of freemium	Assess the feasibility of algorithmic churn prediction in freemium apps; assess the viability of free premium goods as a retention incentive	Detect valuable users early after their adoption of an app	Evaluate different pricing tactics for a promotion targeted to new app adopters
Methods	Large-scale field experiment; three treatment conditions with different onset times of price promotions and a holdout condition without promotions	Benchmarking of four supervised learning algorithms' ability to predict churners; field experiment with two treatment conditions (algorithmic targeting, heuristic targeting) and a holdout condition	Benchmarking of three supervised learning algorithms in the prediction of future premium demand; synthetic oversampling of behavioral data	Field experimentation, bandit methods; practitioner survey
Data	Representative freemium gaming app published on Google's app platform. New users are randomized into four treatment conditions. 160,582 users who complete level one are used in analysis to improve statistical efficiency.	Two representative freemium gaming apps published on Facebook's and Apple's app platforms. Samples of high-value users (N = 10,736 and N = 7,709) to reduce behavioral heterogeneity and to focus on the users with highest value to the firm.	Representative freemium gaming app published on Apple's app platform; sample of 197,665 new app users, observed for a year in the app.	Representative freemium gaming app published on Apple's and Google's app platforms. Six "in-vivo" field experiments with new app users: N1 = 363,440, N2 = 72,243, N3 = 242,604, N4 = 176,222, N5 = 104,052, N6 = 100,821. Survey with 54 freemium practitioners.
Key findings	Regular price promotions lead to an increase in long-term primary demand in freemium online games, without signs of cannibalizing effects of current low prices on future demand	Neural network predictor achieves highest performance across both apps. Close to 90% of churners can be identified for a false positive rate of 20% (AUC=0.93). Predictive churn management improves effectiveness of communication with users manifold; free premium goods do not seem to be a viable incentive to retain high-value churners.	Meta data available at app download (geolocation and device information) allow to identify premium users approximately twice as well as random selection. Adding a week of behavioral data, observed after users' download of an app, increases prediction performance by another 40%. A neural network on synthetically oversampled data achieves overall best prediction performance.	Universal low-price approach (current managerial practice) does not lead to highest profit. Higher profit can be obtained via a personalized skimming tactic that starts select user segments off at price points twice as high as highest managerially recommended price point, and then drops prices following personalized price paths.
Comments	Research collaboration with Stanford University; available as Stanford Graduate School of Business Working Paper No. 3769 (Runge et al. 2019)	Research collaboration with the Artificial Intelligence Lab of Ecole Polytechnique Fédérale de Lausanne; working paper published at the Computational Intelligence and Games (CIG) 2014 conference (Runge et al. 2014); runner-up best paper award	Research collaboration with the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS); earlier version, focusing on synthetic oversampling, published at the Hawaii International Conference on System Sciences (HICSS) 2018 (Sifa et al. 2018)	Presented at the Theory and Practice of Marketing (TPM) conference 2019 at Columbia University, the Uber Behavioral Science Symposium in San Francisco and the Algorithm Hour of Stitch Fix, also in San Francisco

In the setting of freemium mobile gaming apps, a number of conceptual arguments support the postulated importance of promotions:

1. The consumption of digital content, especially online play, is strongly habit-forming (Eyal 2014): Several sources assert that online games spur strong consumption habits (Chen and Leung 2016; Kwon et al. 2016; Nevskaya and Albuquerque 2019; Jo et al. 2020). A long stream of economic and marketing literature posits that promotional approaches can sustainably increase demand for addictive goods (Becker and Murphy 1988; Becker et al. 1991; Katz and Lavack 2002; Chen et al. 2009; Gordon and Sun 2015). In this framing, promotions are helpful in the sale of freemium goods as they can serve as a gateway to more intense use and purchasing.
2. Zero-price inertia: While freemium's zero-price point effectively attracts users to adopt a product or service, e.g., to download an app, it has been shown that consumers exhibit strong inertia around a zero-price (Shampanier et al. 2007; often termed "penny gap," e.g. Carter 2019). An attractive promotional offer is likely to be more effective in removing users from free use of the app and enticing them to spend money.
3. Reduced time to process information and increased search cost: The attention a product can garner is an essential factor in consumers' decision to purchase it (Bettman 1979; Chandon et al. 2009). Sessions in freemium apps are short; Gameanalytics (2019) reports that users spend an average of ten minutes (median: six) in a mobile game before moving on to another activity or app (Yeykelis et al. 2014, 2018; De Haan et al. 2018). Additionally, users may not be aware of premium upgrade options and mobile phone screens are small leading to increased search cost (Ghose et al. 2013). This combination of short attention spans and increased search cost calls for the use of promotions to direct consumers' attention to premium upgrades and their purchase (Bemmaor and Mouchoux 1991).

4. Uncertainty about quality: Freemium games monetize by selling virtual goods such as boosts, avatars or virtual currency in premium upgrades (Lehdonvirta 2009). Virtual goods and in-app purchases more generally are a new product category that may be unfamiliar to consumers (Hamari and Keronen 2017), a promotional approach can hence entice users to try out this new product category and reduce their uncertainty about it (Foubert and Gijsbrechts 2016).

Speaking to this reasoning, the first paper of this thesis (with Jonathan Levav and Harikesh Nair)[1] presents evidence from a large-scale field experiment that randomizes close to one million new app adopters in three promotional treatment conditions and a holdout group without any promotions. Results show that regular price promotions can substantially increase primary demand in the freemium setting of online games. The authors allude to a particular complementarity between (free) use of the product and the premium experience and users' related inability to sufficiently control their consumption that can help explain this strong increase in longer-term demand without signs of cannibalizing effects on future revenue from current low prices. Evidence further does not suggest that regular promotions serve as an adverse quality signal. These findings are novel in light of a long stream of literature cautioning against the use of low-price approaches due to adverse long-term consequences (Mela et al. 1997; Jedidi et al. 1999; Anderson and Simester 2004; Günter and Klapper 2007). It is further foundational to the validation and development of the main thesis of this dissertation.

The second paper (with Peng Gao, Florent Garcin and Boi Faltings)[2] builds on the first paper's findings by investigating the effectiveness of a free promotional bundle targeted to high-value freemium users at risk of disengaging (Ascarza 2018;

[1] An earlier but highly similar version of the paper is available as a Stanford Graduate School of Business Working Paper which this work references as Runge et al. (2019).

[2] An earlier version of this paper was presented at and published in the proceedings of the Computational Intelligence and Games (CIG) conference 2014. It won the conference's runner-up best paper award. The present work references that earlier version of the paper as Runge et al. (2014).

Ascarza et al. 2018). The aim of targeting the promotion is to increase users' retention by affording them a free premium experience that they have a preference for based on their purchasing history in the app. Several algorithms are compared in their ability to identify users at risk of churning in two freemium gaming apps, and the best performing algorithm is then used in an online experiment that targets a promotional offer to high-value potential churners in one of the apps. While the predictive system is effective in reaching users who are more responsive to the firm's outreach effort, the promotion is not able to meaningfully increase users' retention with the product (Ascarza et al. 2016). It seems that a promotion at such a late stage in users' lifecycles may be unable to change their choices pertaining to the focal product as users have irreversibly lost interest in it. Cross-promotion to other products in the firm's portfolio or to other companies (through advertising) may be superior treatments, and promotions in the focal product may need to be targeted earlier in users' lifecycles.

Speaking to the second paper's concluding conjecture, the third paper (with Rafet Sifa and Christian Bauckhage)[3] assesses if the firm can use the data commonly available to firms in this setting to identify high-value users early after their adoption of the app: The authors apply various learners, particularly emphasizing the benefits of neural networks (West et al. 1997) and synthetic oversampling (Weiss 2004), to predict future premium demand of new app adopters. While far from perfect, device, geolocation and early app use information associate significantly with different expectations of users' spending on premium goods, suggesting that different segments of recent app adopters (as identified from available data) may be receptive to different promotional offers due to heterogeneous valuations for premium experiences (Rossi et al. 1996; Acquisti and Varian 2005). The neural network predictor further achieves best performance in comparison to random forest and linear regres-

[3]An earlier version of this paper, focusing on the implementation and benefits of synthetic oversampling, was presented at and published in the proceedings of the Hawaii International Conference on System Sciences (HICSS) 2018. It is referenced here as Sifa et al. (2018).

sion. It can be applied by firms to support decisions of their marketing managers, for example in customer acquisition (Blattberg and Deighton 1996; Seufert 2013).

The fourth paper builds on the proposition that early app data associate with users' expected spending and evaluates different pricing tactics for a promotion targeted to new app adopters. It focuses on profitably personalizing such tactics using the available data and bandit-based experimentation (Li et al. 2010; Schwartz et al. 2017; Bietti et al. 2018; Misra et al. 2019). Results obtained in six large-scale field experiments suggest that price and promotion personalization in freemium settings can be highly profitable, and that it has potential to increase engagement with content through increased access to premium upgrades. The paper further presents results from a survey with 54 freemium practitioners that indicates that current managerial practice excessively favors low-price approaches and likely harms realized profits.

Analyses in different apps and in collaboration with different companies show that firms not using promotions to market their freemium apps are likely to forego 20 or more percent in revenue (which translates to profit as marginal cost to produce and distribute freemium virtual goods is zero – Anderson 2009; Lambrecht et al. 2014), substantiating that promotions and their targeting are indeed essential to the marketing of freemium goods; in particular to support revenue generation from users with lower, and safeguard the beliefs and monetization behaviors of users with higher willingness-to-pay. While analyses take the perspective of a profit-maximizing firm, wider societal implications are also considered, especially in the fourth paper that speaks to issues of data-driven optimization and fairness. The paper devises a price personalization approach that increases purchases of premium experiences by 26%. While this result is highly desirable to the firm, it does not solely benefit the firm but also consumers who would not have obtained access to premium experiences without the personalization. Price discrimination will usually tend to charge higher prices to consumers with higher willingness-to-pay. To the extent that such higher

willingness-to-pay derives from higher income and wealth, price discrimination can have a desirable redistributive effect. In the case of online games, high willingness-to-pay may however often derive from addictive tendencies (Kwon et al. 2016; Nevskaya and Albuquerque 2019; Jo et al. 2020) rather than be reflective of personal wealth. Proactive regulation of data-driven price personalization and optimization could hence be well advised in this setting, particularly to ensure healthy habits in the consumption of online games (Hahn et al. 2010; Nevskaya and Albuquerque 2019).

Summarizing, this dissertation presents novel, rigorous and generalizable applications of advanced analytics to the promotion of freemium goods, providing guidance to firms how they can direct, optimize, and automate their decision making to more successfully achieve their organizational goals in this setting. In particular, the work shows how firms can more profitably target and price promotions of freemium goods in mobile games. Its findings can be expected to generalize to gamified apps such as Tinder and freemium offerings more widely, e.g., to the promotion of subscriptions on a news website or in a music streaming app. The work further contributes to literature on pricing (Pigou 2017; Shapiro 1983; Acquisti and Varian 2005; Nair 2007; Dubé and Misra 2017; Dubé et al. 2017a; Shiller 2020), promotion (Mela et al. 1997; Jedidi et al. 1999; Anderson and Simester 2004; Günter and Klapper 2007), churn prediction and prevention (Ascarza et al. 2016; Ascarza 2018; Ascarza et al. 2018) and algorithmic demand forecasting (West et al. 1997; Berger and Nasr 1998; Fader et al. 2005). The individual papers discuss these contributions in more detail.

Chapter 2

Price Promotions in Freemium Settings[1]

Julian Runge

Jonathan Levav (Stanford University)

Harikesh Nair (Stanford University)

Abstract

The freemium pricing model for digital goods involves selling a base version of the product for free, and making premium product features available to users only on payment. The success of the model is predicated on the ability to profitably convert free users to paying ones. Price promotions (or "sales") are often used in freemium to induce the conversion. However, the causal effect of exposing consumers to such intertemporal price variation is unclear. While sales can generate beneficial short-run conversion, they may be harmful in the long-run if consumers intertemporally substitute purchases to periods with low prices, or use them as signals of low product quality. These long-run concerns may be accentuated in freemium, where the base version is sold for free, so that sales form extreme price cuts on the overall product combination. We work with the seller of a free-to-play gaming app to randomize entering cohorts of users into treatment and control conditions in which promotions for in-app purchases are turned on or off. We observe complete user behavior for half a year, including purchases and consumption of in-game premium goods, which – in contrast to much of the extant literature – enables us to assess possible substitution over time in consumption directly. We find that conversion and revenue improve in the treatment group; and detect *no evidence* of harmful intertemporal substitution or negative inferences about quality from exposure to price variation, suggesting that promotions are profitable. We conjecture that the zero price of the base product that makes its consumption virtually costless, combined with the complementarity between the base product and premium features can help explain this. To the extent that this holds across freemium contexts, the positive effects of promotions documented here will hold more generally.

[1] An earlier but similar version of this chapter is available as a Stanford Graduate School of Business Working Paper referenced here as Runge et al. (2019).

2.1 Introduction

Freemium is a popular pricing model for digital goods (Shapiro and Varian 1998; Shampanier et al. 2007; Kumar 2014; Lambrecht et al. 2014; Lee et al. 2017). In it, firms offer a version of their product or service for free to acquire and engage with consumers, and then upsell premium upgrades that require payment. Examples include: news content (The New York Times), music (Spotify, SoundCloud), file storage and collaboration (Dropbox, Slack), communication (Skype), dating and networking (Tinder, LinkedIn) and digital games (Candy Crush Saga, Farmville). Pricing of premium features takes two common forms. In simple, two-tiered subscription-based freemium, consumers pay in installments to maintain access to a full bundle of premium features. For example, in Spotify, a monthly payment removes advertising and allows for multi-device usage. In Dropbox, upgrading to a premium plan provides expanded storage space. In the New York Times, subscribing to the paid digital version removes the limit on articles and allows customization. In more complex multi-tiered versions of freemium, increasingly attractive tiers of premium features are made available to users as they pay more. Most freemium games (commonly termed "free-to-play" games) use this model: Users pay by purchasing in-game goods in variable quantities, which can be used to unlock a desired level of upgrades or to proceed to more advanced levels of the game. This paper pertains to this type of freemium.

Pricing in a freemium model involves complex issues of how to designate product features into paid and unpaid sets; how to set the prices of the features; and whether and how to dynamically adjust those prices over time. Empirical work on freemium pricing is limited, and many of the key issues are still not well understood in the context of digital goods. In this paper we investigate one aspect of freemium pricing: the effect of dynamic pricing over time. We implement a field experiment for a digital freemium product − a free-to-play gaming app − and investigate the causal effects

for the seller of introducing price variation in the form of periodic promotions or "sales."

This question is interesting in freemium settings, as the success of the freemium model relies on the efficiency with which free users are converted to paying ones. Sales help conversion by bridging the so-called "penny gap" in freemium − a colloquial term to describe a commonly held belief in the start-up community that it is harder to convince a customer to pay the first penny than to induce him to pay more once initial payment has been made (Anderson 2009; Shmilovici 2011; Carter 2019). The difficulty of converting free users to paying ones could arise from a fixed cost to the user of setting up payment (e.g., the customer has to enter his credit card details and get verified) or by the special significance of a "zero price" in consumers' minds (Heyman and Ariely 2004; Shampanier et al. 2007; Ascarza et al. 2012). Apart from improving free to pay conversion, sales can also increase repeat purchases of in-game goods, facilitating unlocking of premium features that improve user experience and increase future retention and gameplay. As such, they are common in free-to-play games, and the conventional wisdom is that sales are beneficial for the seller; a decrease in price increases demand.[2]

On the other hand, there are some reasons to question the conventional wisdom. While short-run conversion is likely increased during promotions, the long-term consequences of periodic sales is far from obvious. The concern is that in situations with repeat purchase and in which consumers have price knowledge, systemic price variation can cause consumer behavior to adjust so as to "game the system." For instance, consumers who anticipate the promotion cycle can time their purchases, delaying current purchases and pulling forward future purchases to low price sales periods. Sales can thus become expensive giveaways, essentially serving to generate unprofitable intertemporal demand substitution, moving a purchase that would otherwise

[2]In-game purchases in free-to-play games accounted for 82% of worldwide digital games revenue in 2017 (Gough 2018). The market for mobile in-game consumer spending alone makes up close to 80% of $101 billion in-app revenue in 2018 (App Annie 2018).

have occurred at a high-price future or past to a low-price present featuring a sale. Indeed, for this reason a body of work in the academic literature has warned that the conventional wisdom about the advantage of promotions must be evaluated with caution (Mela et al. 1997; Nijs et al. 2001; Erdem et al. 2003; Hendel and Nevo 2003; Anderson and Simester 2004; Neslin and van Heerde 2009; Anderson and Simester 2010; Elberg et al. 2019; Nair et al. 2017). Further, in situations where products are "experience goods" and consumers are uncertain about their match-value with the product, a separate literature has warned that consumers can use low prices as a signal of low quality (Gerstner 1985; Milgrom and Roberts 1986; Rao and Monroe 1989; Erdem et al. 2008; Dubé et al. 2017b). If this occurs, exposure to sales can reduce user satisfaction and product usage, harming the seller. The concern about sales signaling quality may be accentuated in a freemium setting where the base version is already being provided for free, so additional price cuts on premium features may be viewed as extreme price cuts on the product.

The research that casts doubt on the long-term value of sales draws its conclusions from studies of non-digital storable, consumer packaged goods (e.g., paper towels, potato chips, coffee, etc.). The purchase of such goods is typically separated from their consumption – a consumer that purchases three bags of chips on sale is highly unlikely to eat them all during the store visit – so that stocking up and purchase planning are a viable alternative response to a sale. In addition, the quality of these goods' consumption is unaffected by the decision to purchase multiple units – the quality of a potato chip is invariant to the number of bags of chips purchased.

In contrast, digital freemium goods differ from typical durable or consumer packaged goods in several respects. First, the base product and premium features tend to be *complements* to each other. Higher consumption of the base product makes it more likely that premium features are added-on; and higher consumption of premium features likely makes the base product more attractive. Second, consumption of the base product is available to the user *for free* given its zero price. Upgrades to

digital freemium products are ordinarily executed in the course of the base product's consumption, as a result of an immediate need to enhance the base product or due to an impulse, which may make immediate consumption of the purchased add-on more likely. These aspects may conspire sufficiently to make saving currently purchased premium features for the future less likely or unattractive. Consequently, the interplay of price variation and stockpiling behavior that occurs in digital freemium goods is an empirical question.

We collaborate with a video game company to implement a field experiment to assess the impact more formally. The game chosen for the field experiment combines a puzzle with a city building component and is representative of the free-to-play game genre. It is published on Google's, Apple's and Facebook's app marketplaces and has been downloaded by more than 50 million players as of October 2017, when the experiment ended. The base version of the game is free to play to all users. Users can buy in-game currency using real money, which they can use to unlock a variety of game features. The question of pricing pertains to the exchange rate between in-game currency and real money. Price variation is induced when the exchange rate for in-game currency is discounted via periodic promotions. Starting mid-December 2016, we randomize cohorts of consumers who download the game app, into treatment and control conditions in which their exposure to price variation is randomly switched on or off. The treated group is exposed to a fixed schedule of periodic promotions. The schedule features "Hi-Lo" pricing analogous to sales in CPG settings (Hoch et al. 1994; Bell and Lattin 1998; Ho et al. 1998; Ellickson and Misra 2008; Ellickson et al. 2012 – specific details are described later in the paper). The control group sees a constant price with no promotions. Individual-level data on game usage and spending for the users are tracked for half a year (180 days) post app download.

Analyzing the data, we find that exposure to the promotion sequence generates large benefits to the firm: Relative to the control group, conversion increases by

37.1%; purchases of in-game currencies increase by 27.2%; and revenue from premium features increase by 23.6%. We detect no evidence of harmful intertemporal substitution of demand for in-game currency: Average outcomes in pre-and post promotion cycles are statistically indistinguishable between the treated and control groups. Thus, it appears that almost all the lift observed during promotion cycles represents incremental expansion of demand and increased consumption of purchased goods. We also see little evidence of prices adversely signaling quality: Login behavior of users remains unchanged between the treated and control group. We further observe usage and spending in two other games in the company's portfolio for a subset of users – which are not different between treated and control conditions – suggesting that regular promotions indeed lead to a substantial increase in *primary* demand in this setting. Finally, we observe a small and statistically only marginally significant decrease in usage in the promotional treatment condition. This result indicates that promotions do not serve as a strong quality signal that would lead users to abandon the app. The authors propose that the significant increase in demand from regular price promotions derives from the strength and immediacy of the complementarity of the free and premium version in freemium products: As users cash in on the deal offered in a promotion, i.e., use the purchased premium feature, their utility derived from time spent using the product experiences a boost. The immediacy of this gain in utility may habituate them towards future purchases, reducing users' ability to regulate their consumption and effectively lowering cannibalizing effects on future purchasing.

To the extent that this complementarity between the base and premium product and the immediacy of utility increases from deal purchases apply in other freemium environments, price variation in the form of sales will be a profitable policy more widely. The complementarity plausibly exists in many freemium contexts. For instance, in Dropbox, additional collaboration opportunities (a premium feature) may make the storage space available in the base version more valuable, and vice

versa, albeit in a more paced manner. Immediate increases in utility from deal purchases seem plausible for gamified products such as Tinder and strongly habit-forming products more generally, possibly reducing consumers' ability to regulate their consumption.

Relationship to the literature. On the data side, the digital app-based environment facilitates some novel aspects of this study: (a) inducing randomization and controlling experiences at the user-level, which improves statistical power, and facilitates exploration of heterogeneity in user-level response; (b) sustaining a control group with no promotions for a long period of time (180 days), which has typically been difficult in many settings due to the costs of such experimentation and the inability to closely control user experiences; and (c) observation of consumption and purchases over time at the user-level, which directly facilitates users' substitution of goods over time. While past papers have leveraged these aspects in isolation, none have brought all three aspects together in a study of pricing to our knowledge. To the best of our knowledge, this is also the first paper that has presented an evaluation of sales in a digital, freemium setting using a randomized controlled trial. Further, we observe consumption, enabling a direct assessment of intertemporal demand substitution. This direct assessment is novel compared to the past literature on storable goods, which has typically observed purchases but not consumption, and has therefore relied on indirect assessments. The paper is related to a subset of papers that randomize users cross-sectionally into different prices so as to study static price discrimination, e.g., Levitt et al. (2016); Sahni et al. (2016); Dubé et al. (2017a); Dubé and Misra (2017). Levitt et al. (2016), which involves virtual goods in a video game and is closely related to this work, has this flavor. Our study is distinct from this stream of papers as it focuses on measuring the causal effect of sustained exposure to a *sequence* of prices over time to investigate its intertemporal consequences for initial and repeat purchases. This paper is related to

experimental studies by Hoch et al. (1994) and Elberg et al. (2019) who implement store- or category-level randomization of pricing policies to study promotions in grocery retail (comparing "Every Day Low Pricing" versus "Hi-Lo" promotions in Hoch et al. (1994) and "Hi-Lo" promotions with deep versus shallow discounts in Elberg et al. (2019)). Given the grocery store setting, these studies do not implement a control group with "no promotions," and are not specifically focused on the interaction of promotions with a tiered system of goods. In contrast, this study relates to digital, freemium goods that comprise base and premium tiers; involves a control group with no promotions (and hence is able to benchmark against a no-promotion environment); and implements randomization at the individual level (which yields individual-level data and more statistical power). Also closely related are field experiments by Anderson and Simester (2004) and Anderson and Simester (2010), which compare catalog purchases between individuals randomized into receiving a catalog with deep or shallow discounts. Anderson and Simester (2004) and Anderson and Simester (2010) find positive long-run impact from exposure to deep discounts for new consumers, consistent with our findings for new users in the studied environment. Our study is distinguished from their work by its focus on freemium; by having access to a control condition with no discounts; and by its focus on studying the treatment effect of a pricing policy that comprise a sequence of varying prices over time, rather than a one-time reduction.

2.2 Empirical Setting and Field Experiment

2.2.1 Empirical setting

The video game on which the experiment is implemented can be described as involving "solving puzzles and building a city."[3] Within the game, the player confronts

[3]See `https://en.wikipedia.org/wiki/Puzzle_video_game` for descriptions of this genre.

graphical puzzles that have to be solved while trying to build a city. The puzzles and the city the player is building are connected to an overarching storyline and metagame. This induces programmatic buildup and continued engagement as gameplay progresses. Free play is possible in perpetuity and the game can be completed without a single purchase. However, if the player chooses to, he can purchase a variety of in-game premium goods to enhance gameplay. Examples include additional energy to extend a game session or decorations that beautify the city the player is building. To purchase in-game goods, the player exchanges real money for two types of in-game currencies. The in-game currencies are sold within the game separately or in bundles. The in-game currencies do not expire, and are exchangeable for in-game goods at rates determined by the game seller.

Users are also sometimes shown in-game ads, typically video-ads by other third-party apps. Users who watch the video-ads are rewarded a small amount of in-game currency. On seeing the ad, they are shown a still-screen with a link to the app store.

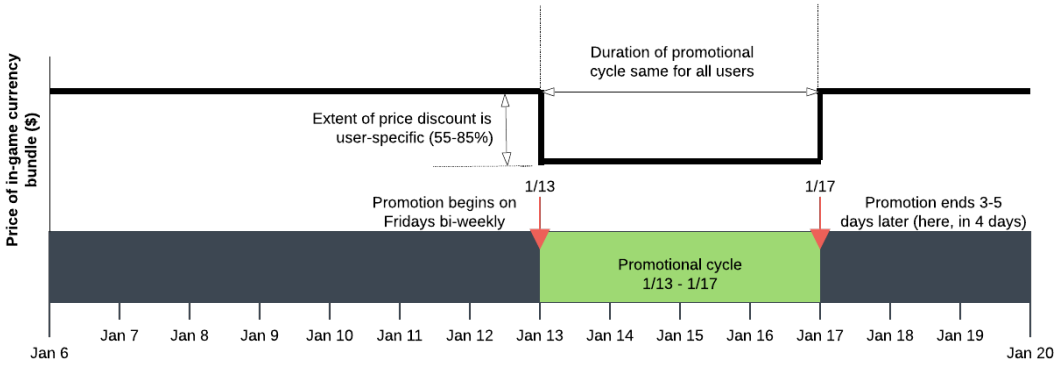
Advertising makes up only a small percentage of the revenue from the product ($<$10%); the bulk of the revenue is from purchase of in-game currency. This monetization model is common in the free-to-play gaming industry (see Figure 2.10 on page 37 in the Appendix for an example).[4]

2.2.2 Experimental design

To implement the experiment, we isolated a sub-population of new users who downloaded the game application from Google's Playstore to their Android smartphone or tablet between December 15, 2016 and February 21, 2017. On download, we randomized them to a treatment group and a control group as follows.

[4]This is also representative of the "app economy" where advertising only drives a small share of revenue (Ghose and Han 2014).

Figure 2.1: Promotion schedule in treatment group: example for 14-day period from January 6, 2017 to January 20, 2017



Notes: The promotion sequence that treated users are exposed to involves "Hi-Lo" pricing: In non-promotional periods, the in-game currency is sold at a fixed base price. Every second Friday, a discounted bundle is shown to the user. The specific bundle picked depends on the user's past buying history. The discount offer is also personalized to the user – it ranges from 55% for users who purchased in the past (promotionally or not) to 85% for users who never made a past purchase. The promotion lasts for 3-5 days. A user can only make one promotional purchase per cycle, and the length of the promotional cycle is fixed across all users.

Treatment group:

- Each user in the treatment group is exposed to a *sequence* of promotions for half a year (180 days) from the time she downloads the game.
- The promotions pertain to a bundle of in-game currencies offered to the user within the game.
- The promotion plan is as follows: In non-promotional periods, all in-game currencies are sold at a fixed base prices in a game store accessible within the video game. Every second Friday, the user is shown a specific in-game currency *bundle*. The bundle represents a discount on the component in-game currencies relative to buying them separately in the game store. The specific bundle the user is shown, and the extent to which it is discounted during the sale is personalized to the user (on the basis of the recency, frequency and monetary value of his past purchases). Broadly speaking, larger currency bundles are shown to users with more past purchases and larger percentage

discounts are offered to non-buyers than buyers. The bundle prices range from $1.99 to $99.99, and the discounts range from 55% for users who purchased in the past (promotionally or not) to 85% for users who never made a past purchase. The discounting lasts for three to five days. A user can only make one promotional purchase per cycle, and the length of the promotional cycle is common across all users. Apart from offering the bundle, no other price reductions of the in-game currencies are offered.

- When the promotion is active, it is advertised to the user with an in-app "pop-up." The user can choose to purchase the bundle offer by clicking on the pop-up or send it to the background. If sent to the background, it will remain as an icon on the game app's main screen until purchase or the end of the promotional cycle. Once a purchase occurs or the cycle ends, the icon disappears from the user's screen. Figure (2.1) presents an illustration of the sequence using the 14-day period from January 6, 2017 to January 20, 2017 as an example.

Control group:

- Users in the control group are not exposed to promotions. They see the in-game currencies within the game store at the same fixed base price as the treatment group. Everything else is held the same between the treatment and control groups.

Randomization is persistent: i.e., once a user is allocated into a group, he stays in that group for the next 180 days. The behavioral targeting rule and the exact length of the promotional cycle each active Friday is picked by the firm, and not controlled by us. The firm determined this particular behavioral targeting rule and discounting percentages based on heuristics it developed from past experience, and was unwilling to allow us to vary these aspects.

2.2.2.1 Improving statistical efficiency

User attrition for free-to-play games is high, and a large proportion of users tend to drop off after initial trial. The attrition has the potential to induce significant noise into the statistical analysis, reducing precision.[5] To address this, we built in the following guardrail into the experimental design. The game has 90 levels. User attrition is highest prior to reaching level one. Therefore, we expose users in the treated group to the promotion cycle only after they reach level one in the game. Figure 2.2 shows this pictorially.

In Figure 2.2, a user i in the treated group reaches level one in τ_i days, so is exposed to the treatment for $d_i = 180 - \tau_i$ days. We compare the behavior of all users in the treatment and control groups who have crossed level one. Since users are not exposed to any promotions before they reach level one, the set of users who reach level one in the treatment and control group remain balanced. Figure 2.11 on page 38 in the Appendix presents a quantile-quantile plot of the distribution of d_i in the treatment and control groups and shows they are the same, confirming this.

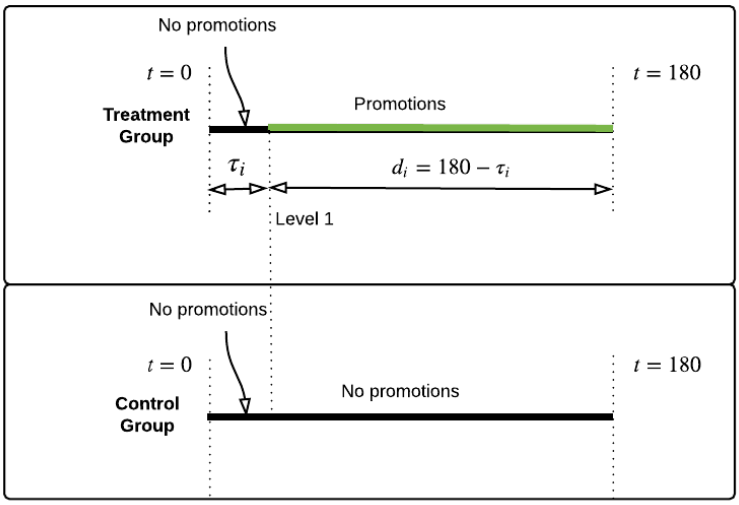
2.2.2.2 Sub-treatment conditions

Due to business and technological constraints, randomization of the *characteristics* of the promotion sequence (e.g., base price, discount depth), was infeasible. Nevertheless, the firm allowed us to vary the timing of exposure of users to promotions. The treatment group described previously is thus created by randomizing users into three sub-groups based on when they start seeing the promotion sequence after download:

1. *Immediate Treatment Group*: This group is exposed to promotions once they cross level one.

[5] A treated user is exposed to the treatment for 180 days. A user who drops off is exposed to the treatment for $\leq$ 180 days. Thus, the intensity of treatment is lesser for the attrited group. We measure an intent-to-treat effect (ITT), comparing all users who were randomized into treated versus control groups (including those that drop off). The noise in the ITT is higher the higher the number of users who drop off.

Figure 2.2: Experimental setup



Notes: To improve statistical efficiency, users in the treatment group have to complete level one before they are exposed to the promotional schedule, and analysis is based on all users who cross level one. Since users are not exposed to any promotions before they reach level one, the set of users who reach level one in the treatment and control group remain balanced. A user i in the treated group reaches level one in τ_i days, so is exposed to the treatment for $d_i = 180 - \tau_i$ days. Figure 2.11 on page 38 in the Appendix shows that the distribution of d_i in the treatment and control groups is balanced.

2. *25-day Delayed Treatment Group*: This group is exposed to promotions once they cross level one *and* it has been at least 25 days since they downloaded the game.
3. *50-day Delayed Treatment Group*: This group is exposed to promotions once they cross level one *and* it has been at least 50 days since they downloaded the game.

Once a user is in a sub-group, he remains in that group for the duration of the experiment.

For most of the initial analysis below, we pool data across the sub-groups to improve statistical efficiency, comparing the treatment group overall, which is exposed to promotions, to the control group which is not. We use comparisons of behavior across the three groups as a way to assess the in how far price serves as a signal of quality. The idea is that under price signaling, users who are exposed to price discounts earlier may have a stronger propensity to infer that the game is of low quality *ceteris paribus*. So evidence of reduced logins to the app when exposure to

promotion is earlier is one indication of signaling effects. We want to emphasize that a more direct test of price signaling would be to randomize the magnitude of the price discounts which, unfortunately, was not possible.

2.3 Analysis

A total 160,582 users reach level one across treatment and control groups and remain after data cleaning.[6] The treatment group sizes were decided in collaboration with the firm who had a managerial prior in favor of higher exposure to price promotions. 26,521 users are assigned to the control and 134,061 to the treatment conditions, with 54,883 users allotted to the "immediate" treatment sub-group; 52,652 to the "25-day delayed" treatment sub-group; and 26,526 to the "50-day delayed" treatment sub-group. Tables 2.3 and 2.4 on page 39 in the Appendix report tests of balance across the groups and show that randomization is induced properly.

2.3.1 Overall treatment effects

Table 2.1 on page 24 reports on the differences in means between treatment and control groups for user free to pay conversion (i.e., indicator for purchase of in-game currency over the duration of the experiment); purchases (total number of orders of in-game currency over the duration of the experiment); revenue from premium upgrades (total money spent on in-game currency over the duration of the experiment); ad revenue and usage (total time spent in app in hours over the duration of the experiment). The main findings are summarized below:

[6]The following filters are applied for data cleaning. Data are observed at the device level. About 50% of devices are connected to one or more Facebook accounts. We disregard all devices that are connected to more than one Facebook account as this indicates that the game was played by more than one user. We further disregard devices that are connected to Facebook accounts that previously played the game on other platforms, and we disregard devices with Facebook accounts that were connected to multiple devices during the experiment. The remaining sample contains devices for which we can reasonably assume that only one user played the game. This data was further cleaned for duplicate and obviously broken log entries and for devices that experienced technical difficulties or where the user "hacked" the app during the experimental period.

- Free to pay conversion increases by 37.1% ($p = 0.000$), from 4.3% in the control group without price promotions to 6% in the promotional treatment group.
- Purchases of and revenue from premium features increases by 27.2% ($p = 0.000$), and 23.6% ($p = 0.026$), respectively, from an average of 0.29 purchases and \$3.53 in the control group, to 0.37 purchases and \$4.36 in the treatment group.
- Advertising revenue falls by – 11.7% ($p = 0.000$) from \$0.29 in the control group to \$0.26 in the treatment group.
- Overall revenue (sum of revenue from premium features and advertising) increases by 20.9% ($p = 0.033$) from \$3.82 in the control group to \$4.62 in the treatment group.
- Game usage (measured as total time spent on the app in hours) falls in the treatment group, but the effect is not statistically significant at the 95% level.
- Use and spending in other portfolio games of the data sponsor are not different between treated and non-treated users. As only about 8% of the 160,582 users considered in experiment analysis used a portfolio game during the experiment period, this result should not be regarded as conclusive.

These results present an unexpectedly favorable picture of promotions. Exposure to promotions improves conversion and revenue, with little detectible negative effects on usage and engagement. Even though advertising revenue falls in response to treatment, advertising revenue only makes up about 8% of overall revenue (using values from the control condition); so overall revenue improves from exposure to treatment due to the increased take-up of in-game purchases.

2.3.2 Assessing intertemporal substitution

To assess the extent of intertemporal substitution, Figure 2.3 displays a time series over the calendar days of the experiment, of the difference in mean per-day revenues and purchases between the treated and control groups. The calendar days corre-

Table 2.1: Comparing user behavior in the treatment and control groups

	Treatment: Promotions N = 134,061 Mean	SD	Control: No promotions N = 26,521 Mean	SD	Difference (T−C)	p-value of Difference
Conversion (indicator for in-game purchase)	0.059	0.236	0.043	0.204	0.016 ***	0.000
Purchases (no. of orders)	0.37	3.20	0.29	3.09	0.078 ***	0.000
Revenue from premium upgrades (A = a1 + a2)	4.33	57.12	3.53	55.32	0.807 **	0.031
Revenue from non-promoted upgrades (a1)	3.20	45.59	3.53	55.32	−0.329	0.363
Revenue from promoted upgrades (a2)	1.14	15.49	0	0	1.14	–
Advertising revenue (B)	0.255	0.845	0.29	0.921	−0.034 ***	0.000
Overall revenue = A + B	4.59	57.12	3.82	55.32	0.773 **	0.039
Usage (total time spent in app in hours)	179.0	409.5	184.3	419.9	−5.27 *	0.061
Revenue in other portfolio games	0.136	10.10	0.121	4.35	−0.015	0.690
Sessions in other portfolio games	6.43	57.94	6.64	66.64	−0.204	0.6412

Notes: The table presents means and standard deviations, intent-to-treat effects and p-values for conversion, purchases, revenue and usage. Statistical significance reported using t-tests with unequal variance for binary outcomes; *** significant at the 1%-level, ** 5%-level, * 10%-level.

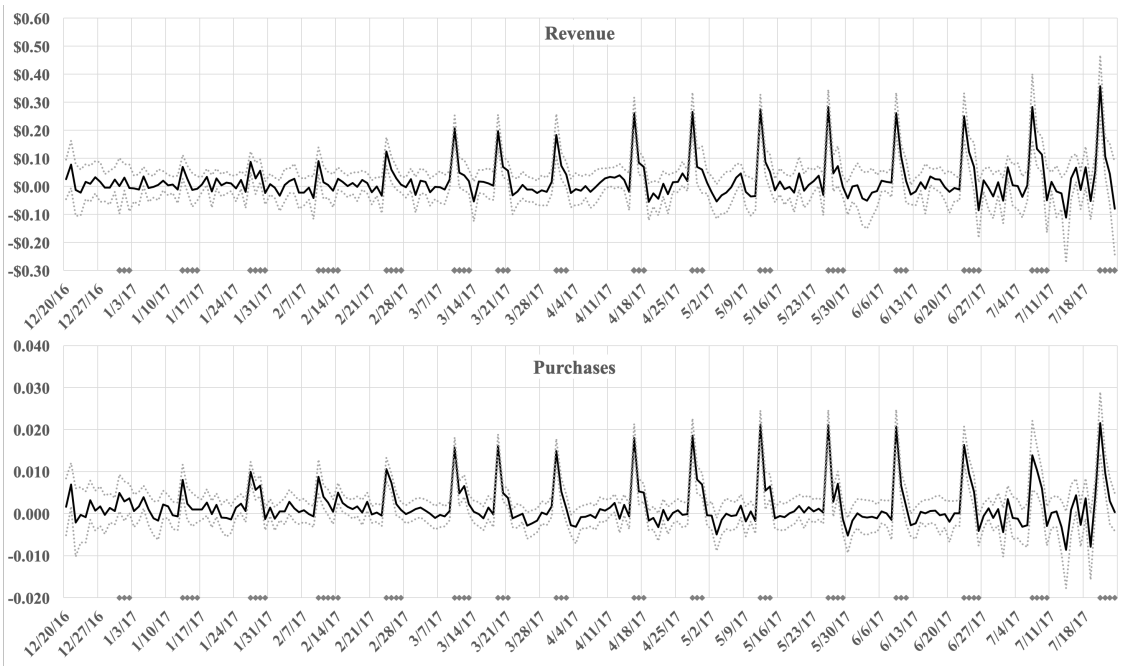
sponding to a promotion cycle are indicated with grey diamonds on the date axis. The dotted lines represent 95% confidence intervals for the difference in means each day. To see the effects visually, we plot this for active users each day, i.e., those who log in to the app on a given day.[7]

Looking at Figure 2.3, we see pronounced regular spikes in both revenue and purchases that coincide with promotional cycles. There is little evidence of reduction prior to a promotion cycle or after one, which would suggest intertemporal substitution. Rather, it seems the spikes reflect expansion in primary demand for in-game currency. Figures 2.14 and 2.15 on page 42 in the Appendix show the same plots separately for the treatment and control groups. Eyeballing these, we can see there is no evidence of systematic demand spikes during the promotional days in the control group, unlike the treatment group.

To hone in on the days before and after promotions, Figure 2.4 pools all the promotion cycles together and displays the mean revenues in treatment and control groups during and around the days of the promotional cycles. The "dashed line" presents the mean in the treatment condition, while the solid line represents the

[7]Figures 2.12 and 2.13 on page 41 in the Appendix document that there is no difference in users' propensity to log in by treatment versus control group.

Figure 2.3: Mean difference in per-day revenues and purchases between the treated and control groups



Notes: Mean difference in revenue and purchases per day with 95% confidence band over calendar days between control and promotional treatment groups. Grey diamonds on the date axis indicate days when promotions were active.

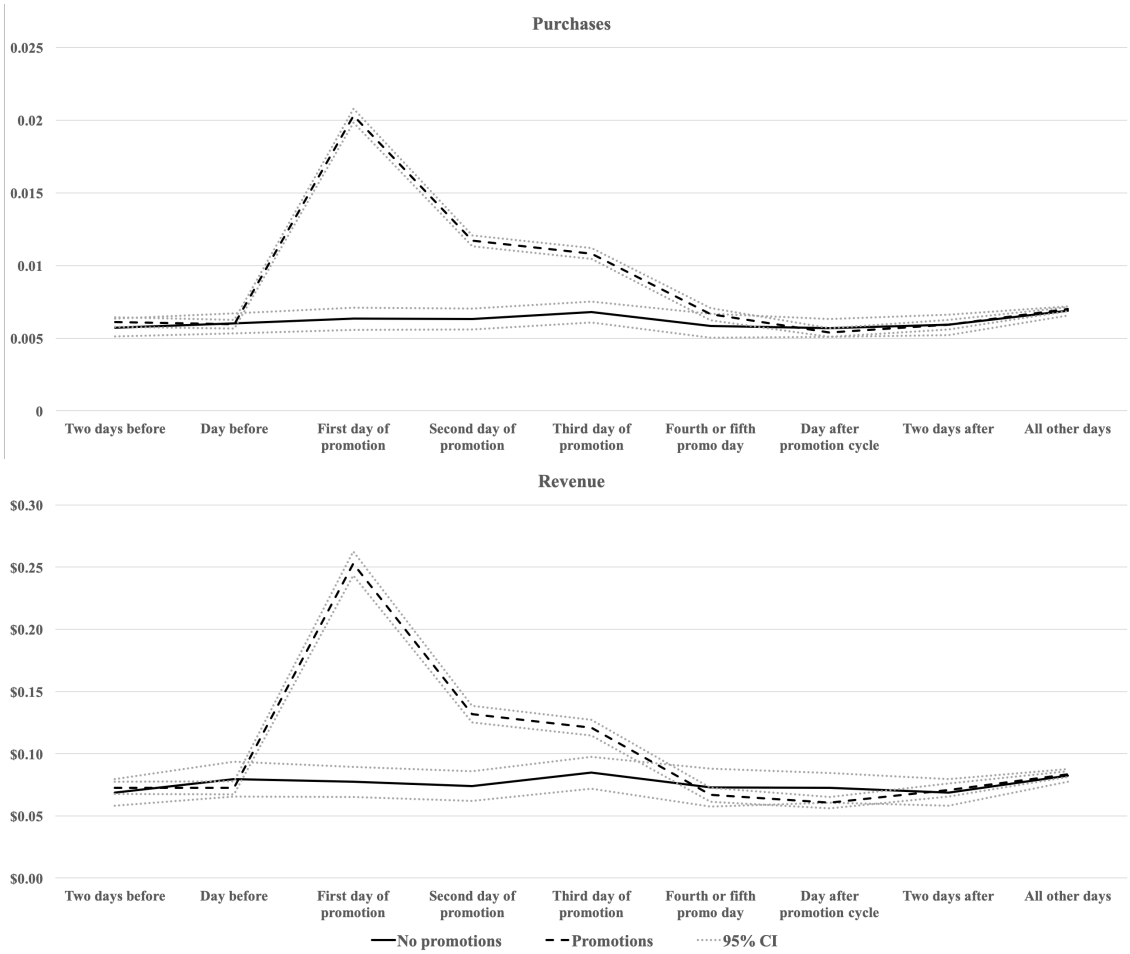
mean in the control condition, and the dotted intervals represent 95% confidence intervals. Looking at Figure 2.4, we see there is little difference between the two groups before or after the promotion cycle, suggesting no detectable intertemporal substitution.

2.3.3 Digging deeper: Why no intertemporal substitution?

What could explain these results? Answering this credibly is difficult as our experiment was not explicitly designed to test between competing mechanisms, so we cannot convincingly rule out one explanation versus another with our data. We offer some conjectures of mechanisms that we believe can plausibly explain our data along with informal empirical support.

One possibility is that two aspects that distinguish digital freemium goods – complementarity between base and premium features and free consumption of the

Figure 2.4: Mean per-day purchases and revenue in the treated and control groups during and around promotional cycles



Notes: Purchases (no. of orders) and revenue (in $) per user per day during and around promotional cycles with 95% confidence band for control (solid line) and promotional treatment (dotted line) groups. More user-day observations are behind the data point for "All other days," hence the confidence band is more narrow there.

base good − make it difficult for users to *regulate their consumption* sufficiently to generate the kind of intertemporal substitution required for sales to be unprofitable. In non-promoted periods, users may find it hard to regulate their consumption of in-game goods (and postpone them to future low-price periods) because the base version is available for free and its consumption enhances the value of the premium upgrades. In promoted periods, users may again find it hard to regulate their consumption of in-game goods (and stockpile for future high-price periods) because the base version is again available for free and its consumption enhances the value of the premium upgrades. Further, consumption of the premium upgrades, which are

then available for low prices, enhances the value of consuming the base good, feeding the cycle. These features are implied by the "addictive" nature of gaming, and may be especially salient for the subset of "hardcore gamers" that pay substantially for games.

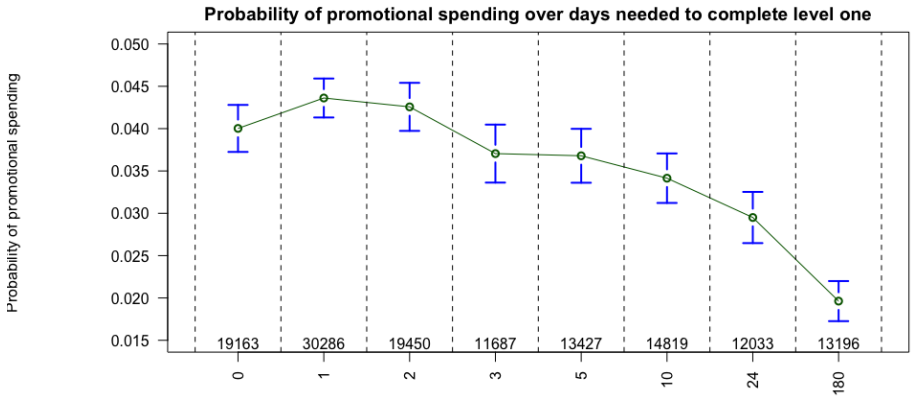
We assess support for users' lack of consumption regulation in our data in two ways. Our first assessment is exploratory, testing whether "heavy users" − who value the game more and may derive more immediate utility from premium feature consumption − are the ones who buy more promoted goods. We proxy for heavy users using the time taken to reach level one in the game. Heavier users are expected to reach level one faster. Since randomization into experimental groups is done only after users reach level one (as explained in more detail below), this forms a valid pre-experimental baseline characteristic on the basis of which to explore subsequent purchase behavior.

Figure 2.5a presents plots of the probability of spending money on a promotion during the experiment as a function of the time taken to reach level one. Figure 2.5b shows the same plot for the amount of money spent on promotions for those who purchased something on promotion during the experiment (i.e., spend|spend$>$ 0). Both are found to be higher for those who reached level one faster.

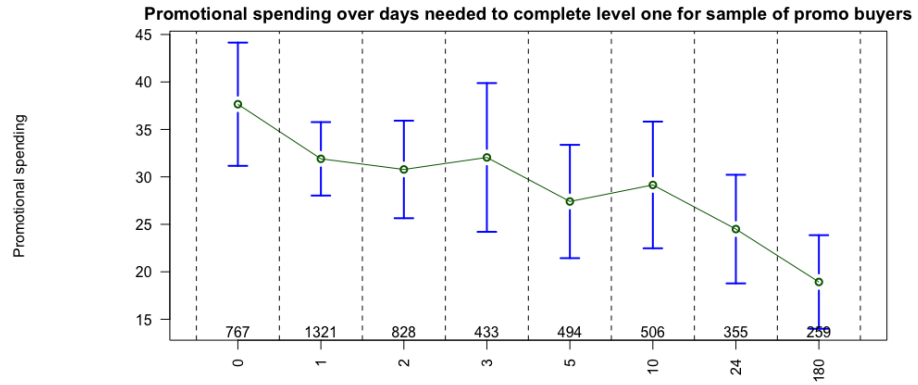
Our second assessment is more direct, and leverages the fact that we can observe consumption directly in our data as we can track the consumption of the in-game currencies purchased by users. To assess self-regulation, we test whether users spend the in-game currency in the same period as they purchase it, or whether spending occurs later. This test is complicated by the way in which in-game currencies are used in the game. The game has three types of in-game currencies: cash, coins and energy. The promotions offered comprise bundles of the three currencies. Cash is more valuable than coins or energy as it can be used to enhance gameplay, advance to higher levels, and also to acquire coins. Also, cash is never earned nor

Figure 2.5: Promotional spending as a function of time needed to reach level one

(a) Extensive Margin



(b) Intensive Margin

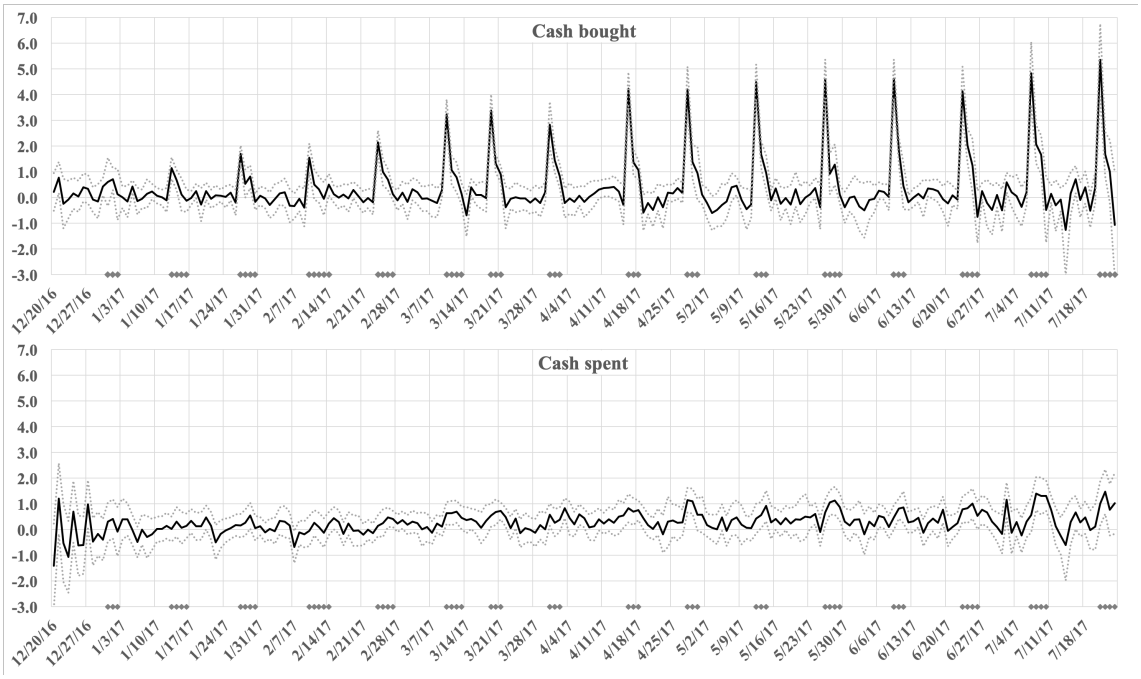


Notes: The figures present the probability of spending money on promotions during the experiment (top), and the amount of money spent on promotions conditional on non-zero promotional spending during the experiment (bottom) on the y-axes, as a function of the time (in days) time to reach level one on the x-axes. Days needed to reach level one are grouped into similarly sized buckets and the x-axis lists the upper inclusive threshold of a bucket. E.g., users in the first bucket reached level one on the day of app download (0), users in the second bucket reached level one within > 0 and $<= 1$ days after app download and users in the last bucket reached level one within > 24 and $<= 180$ days after app download.

replenished by gameplay and *has* to be purchased using real money.[8] We observe the redemptions of cash and coins during gameplay as well as the stock of energy available. The interpretation of the redemption of coins is ambiguous as it is not clear whether the coins redeemed were actually purchased (because coins can also be earned from gameplay); or whether it indirectly reflects the redemption of cash

[8]All users start with an initial endowment of cash, coins and energy. Coins are used to enhance gameplay while energy is required to continue gameplay. Coins and energy can be earned through gameplay and are replenished often (with the difference that coins are earned by "advanced" gameplay, and energy can be earned by "typical" gameplay). Such configurations are typical for the puzzle game genre (e.g., see `https://www.gamesparks.com/blog/looking-at-in-game-currencies/`, accessed February 21, 2019).

Figure 2.6: Difference in cash bought and redeemed between treated and control groups



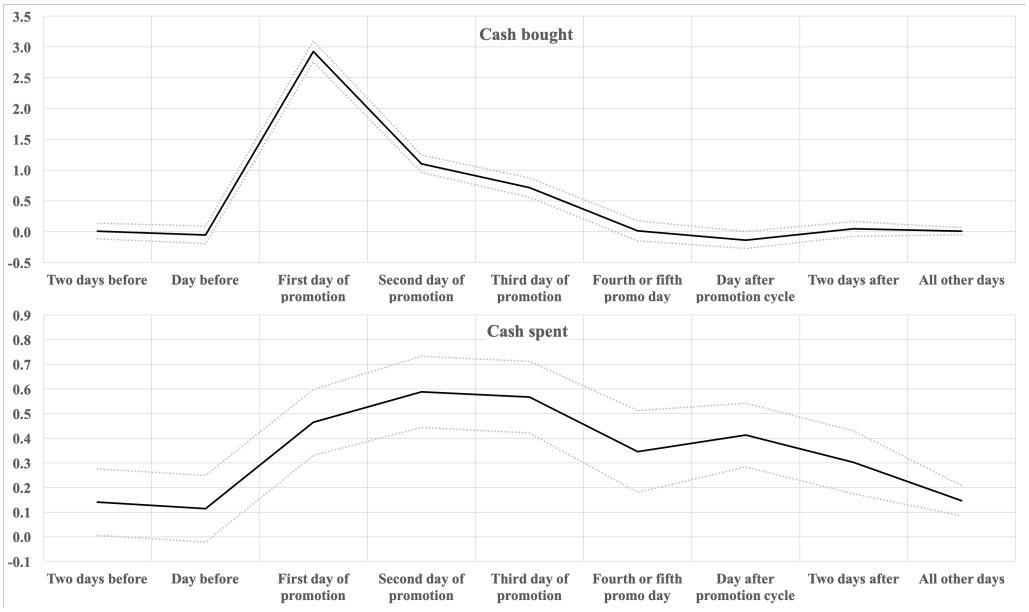
Notes: The figures presents the difference in mean cash bought (top) and cash redeemed (bottom) between experimental groups. Grey diamonds on the date axis indicate days when promotions were active.

(because cash can be used to obtain coins). The interpretation of the redemption of cash is cleaner as additional cash can only be obtained through real-money in-app purchases. Therefore, we focus on patterns of redemption of cash by users.

Figure 2.6 shows the difference in means for the cash bought and cash redeemed between the two experimental groups. Looking at the top panel in Figure 2.6, the cash bought is seen to spike during the promotional periods in the treated group relative to the control. The bottom panel of Figure 2.6 shows a similar pattern in the cash redeemed, though it is less pronounced. We maintain the same scale on the y-axis to make this difference apparent.

To hone in on the days before and after promotions, we create an analogous plot to Figure 2.4 which pools all the promotion cycles together and displays the mean difference between treatment and control groups in cash bought and redeemed during and around the days of the promotional cycles. This is shown in Figure 2.7. Looking at Figure 2.7, we see that cash spent in the treated group rises statistically

Figure 2.7: Difference in cash bought and redeemed between treated and control groups during and around promotional cycles



Notes: The figures presents the difference in mean cash bought (top) and cash redeemed (bottom) between experimental groups, aggregated for days around promotional cycles. More user-day observations are behind the data point for "All other days," hence the confidence band is more narrow there.

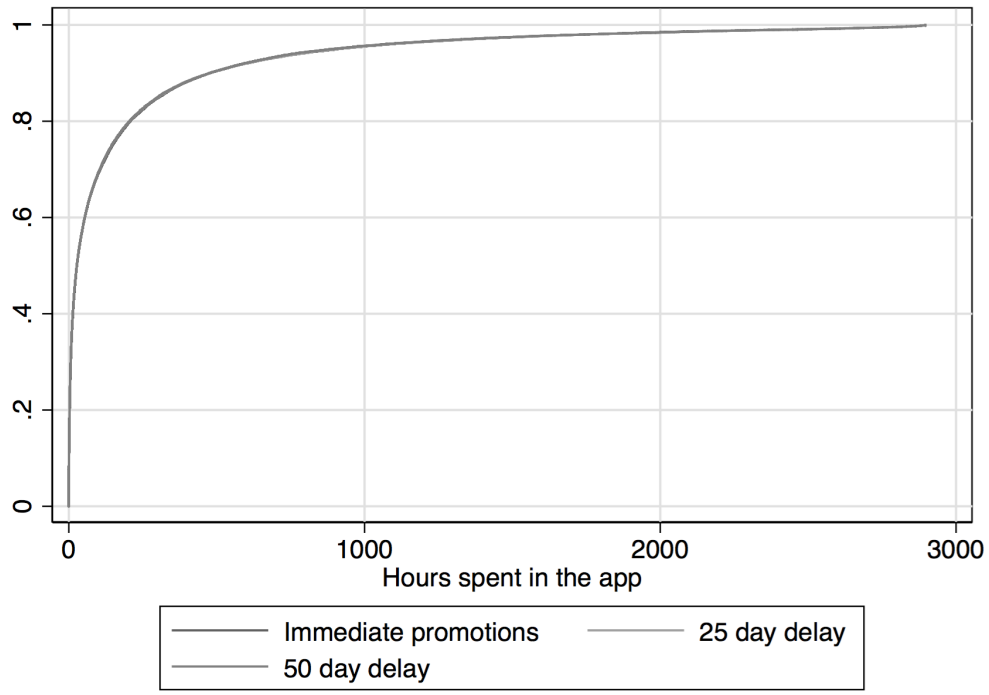
significantly above that in the control group during the days of the promotion cycle, similar to the pattern for cash bought, suggesting that a large subset of users spends the cash bought during promotions almost immediately. The immediate consumption suggests low regulation. Again, we reiterate that we offer these only as simple explanations that can explain our results; testing more formal theories of consumption in freemium is beyond the scope of this paper.

Another possible explanation is that the results simply reflect lack of awareness amongst users about promotions. Towards the end of the paper, we discuss why this may be less likely given the wide publicity of the sales policy of this game on online gaming forums that are frequented by gamers that purchase in-game goods.

2.3.4 Assessing price as a quality signal

To assess whether promotions serve as signals of quality, we compare game usage, measured as time spent in the app. If users perceive the game has lower quality in

Figure 2.8: Empirical CDFs of hours spent in the app by treatment sub-group

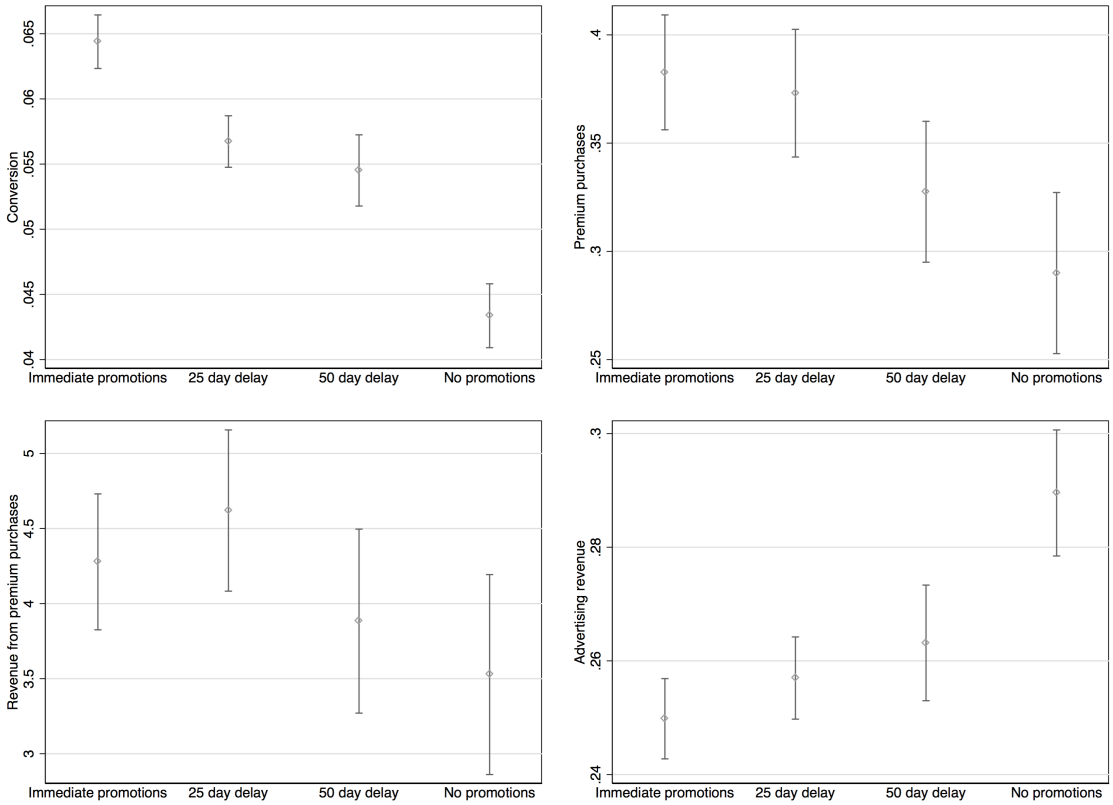


Notes: The plot shows the empirical CDF of total hours spent in the app for the three promotional sub-groups. All three CDFs are plotted, but are visually indistinguishable from each other in the figure.

response to seeing it frequently discounted, we expect they may reduce usage or stop playing the game. This would be reflected in the observed time spent by users in the treatment and control groups. Table 2.1 on page 24 already noted that there is no statistically significant difference in mean time spent between treated and control groups. Figure 2.8 plots the empirical cumulative distribution function (CDF) of the time spent in the app for the three sub-treatments of the treated group, in which the onset time of promotions was varied (as described in Section 2.2.2.2 on page 20). There is no discernible difference in time spent across the three sub-groups. If users who are exposed to promotions sooner and more would take this as a signal of low quality, we would expect their propensity to log into the app and to spend time in the app to go down – neither is the case as shown in Figure 2.8 for time spent and in Figures 2.12 and 2.13 on page 41 in the Appendix for login behavior.

For added confirmation, Figure 2.9 looks at conversion, purchases, revenue and

Figure 2.9: Impact of different promotion onset times on monetization outcomes



Notes: The charts show the mean with 95% confidence interval for key monetization outcomes (noted on the y-axes) across promotional sub-treatments that vary the onset time of price promotions after users' adoption of the app, and for the control condition.

ad revenue across the three sub-groups. Generally, the patterns go in the opposite direction of an adverse quality signaling story: Conversion and purchases over the long-run are higher the earlier promotions begin.[9] Overall therefore, these results do not provide compelling evidence for meaningful harmful consequences of sales to induce negative inferences about game quality.

[9]We also see in Figure 2.9 that advertising revenue is lower when price promotions begin earlier. Given the patterns in Figure 2.8, this is not driven by a mechanical linkage of advertising exposure and advertising revenue to time spent in the app. The fact that early onset of promotions reduces advertising revenue without changing time spent in the app significantly, suggests that promotions and advertising are substitutes from the perspective of users. This can occur for instance, because both provide ways to obtain in-game currencies, or because both compete for scarce user attention. There is some evidence in the literature that promotions can serve as advertising (e.g., Sahni et al. 2016 and the literature cited there), so it is possible that promotions tend to deplete a "mental account" associated with attention to advertising.

2.3.5 Is the positive effect of promotions driven by heavy users?

In terms of treatment effect heterogeneity, we wish to deepen two aspects of the analysis:

1. It could be argued that the surprising effectiveness of regular price promotions is driven by the segment of "hardcore" and possibly addicted gamers.
2. Figure 2.9 shows that, while purchases and conversion are highest in the immediate promotion onset condition, revenue in this condition is lower than in the condition where promotions start with a 25-day delay. It could be hypothesized that this result may be driven by an adverse effect of the firm's behavioral targeting rule that offers a low-price promotion with 85% discount to users who have not yet made a purchase. High-value users complete level one more quickly (see Figure 2.5 on page 28) – they may hence be exposed to the low-price high-discount promotional offer before they have a "chance" to make a possibly larger and non-discounted purchase, adversely impacting their price beliefs and future purchasing behavior.

To assess the merits of these two perspectives, we leverage the fact that device-related exogenous covariates that we observe strongly associate with expected heavy spending on the game. The memory of the user's device and the width of the device screen correlate most strongly with users' overall spending, as observed in the control condition without promotions. A regression of overall revenue per user on both covariates suggests that the width of the device screen more strongly associates with high expected spending in the control group (0.2 $-cents more revenue per additional pixel, $p = 0.0308$, versus 0.03 $-cents more revenue per additional megabyte memory, $p = 0.4669$; N=26,521).

Table 2.2 shows results of a regression of overall revenue per user on treatment indicators and users' device screen width. Model M2.1 confirms results of mean comparisons in Table 2.1 on page 24 and Figure 2.9 on page 32: Overall revenue

is only statistically significantly different from the control condition in the 25-day delay sub-treatment group. Users in this condition spend \$1.06 more within 180 days after app download than users in the control condition without any promotions ($p = 0.0132$). Users in the immediate promotion onset sub-treatment spend \$0.71 more than users in the control in the same time period; this result is only marginally significant with $p = 0.0943$. Both models M2.2 and M2.3 confirm a strong main effect of user device screen width on the degree of spending. Speaking to perspective (2) above, they further indicate that a possibly adverse effect of immediate versus 25-day delayed promotion onset is not driven by a negative impact of immediate promotions on high-value users' spending habits or price beliefs – interaction terms both with continuous screen width in pixels and with a high-low indicator based on a median split are negative. Speaking to perspective (1) above, they further suggest that the strong positive overall effect of regular price promotions is not driven by heavy ("hardcore") users.

We take the assertion that the effectiveness of price promotions in this setting is not driven by a factor specific to games, e.g., addictive consumption patterns (Nevskaya and Albuquerque 2019; Jo et al. 2020), and applies to freemium settings more widely to a further empirical test. We run a regression of overall revenue on a pre-treatment indicator of user engagement, i.e., how fast users completed level one, and a binary treatment indicator. Being in the top 50% of pre-treatment engagement has a strong main effect (+\$1.39, $p = 0.0466$), but the interaction with the treatment indicator is not significant (+\$1.11, $p = 0.2447$), suggesting that the effectiveness of promotions is not disproportionately driven by heavy game users.

2.4 Discussion and Conclusion

Overall, experimental results paint an attractive picture for the use of promotions in freemium settings: Conversion, purchases, and revenue increase with little reduction

Table 2.2: Treatment effect heterogeneity in expected user spending

	Outcome (y): Overall revenue		
	M2.1: Treatment main effects	M2.2: Screen width in pixels (continuous)	M2.3: Screen width indicator based on median split
Control condition excluded as a reference			
Immediate promotion onset	0.7110 *	1.5416	1.1233 *
	(0.0943)	(0.3165)	(0.0421)
25-day delayed promotion onset	1.0600 **	0.5471	1.2364 **
	(0.0132)	(0.7243)	(0.0266)
50-day delayed promotion onset	0.3297	2.5475	0.9762
	(0.5041)	(0.1543)	(0.1283)
Screen width	–	0.0027 ***	2.4935 ***
		(0.0018)	(0.0004)
Interactions			
Immediate * Screen width	–	−0.0006	−0.9765
		(0.5791)	(0.2586)
25-day delay * Screen width	–	0.0004	−0.4380
		(0.7349)	(0.6144)
50-day delay * Screen width	–	−0.0015	−1.5642
		(0.1973)	(0.1191)
Intercept	3.8168 ***	0.0058	2.7919 ***
	(0.0000)	(0.9964)	(0.0000)
N	160,582	160,582	160,582

Notes: Regression of overall revenue per user on treatment indicators and strongest exogenous indicator of expected user spending (screen width of mobile device); *** significant at the 1%-level, ** 5%-level, * 10%-level.

in usage. While advertising revenue is reduced, in this setting, advertising forms a small component of overall revenue, so its economic significance is limited. However, if advertising takes a higher stake in overall monetization, this effect may dominate economically, and therefore, we think this finding is of separate interest.

A simple explanation for the observed lack of intertemporal demand substitution is that users are unaware that promotions will be offered bi-weekly and hence unable to plan ahead. While this is plausible, we find this explanation less compelling for this context. The company has been running bi-weekly sales on the game for a long period prior to the experiment. Engaged players discuss details of the game's economy and sales in online forums (see Figure 2.16 on page 43 in the Appendix). It is hence likely that they are aware of the regular bi-weekly occurrence of sales.

In the introduction, we conjectured that the zero price of the base product that makes its consumption virtually costless, combined with the complementarity be-

tween the base product and premium features and the immediacy of its effect can help explain the behavior. In gaming in particular, such complementarities may be strong due to the "addictive" nature of the gameplay, and this effect may be especially salient for the subset of hardcore gamers that pay substantially for games. While we present evidence that the positive effect of promotions is not driven by hardcore gamers (see Section 2.3.5), and while it is not unreasonable that most freemium base and premium goods are complements, the quantitative significance of such complementarities in other settings is an empirical question. It warrants further, careful empirical testing before further conclusions can be drawn or generalized. Also valuable would be experiments that vary the characteristics of the sales strategy (depth, frequency, items), which would enable studying the empirical consequences of these more deeply and help in the formulation of better policies. The role of addiction and self-control for such goods and their implications for pricing strategy would also be interesting areas of future study to extend current literature considering usage restrictions (Nevskaya and Albuquerque 2019; Jo et al. 2020).

Speaking to the effect of the immediacy of the utility gain from the base and premium complementarity, the results of this study may have normative implications for the design of freemium experiences similar to suggestions by managerial work (Eyal 2014; Alter 2017): Managers can create and reinforce the conjectured complementarity by ensuring that users experience a strong increase in utility immediately after they make a premium purchase. This can be achieved, e.g., by progressing users substantially through a level (games), matching them with much better mates (dating) or professional connections (networking), or substantially increasing the curation and recommendation quality of content (music, movies, news) immediately after a purchase. The stronger the complementarity between free and premium experience and the more immediate the gain in utility from adoption of premium content, the more effectively can regular promotional activity be expected to increase overall demand.

2.5 Appendix

In-game currency and sales in Candy Crush Saga

Figure 2.10: Examples of in-game currency and sales in the popular video game Candy Crush Saga

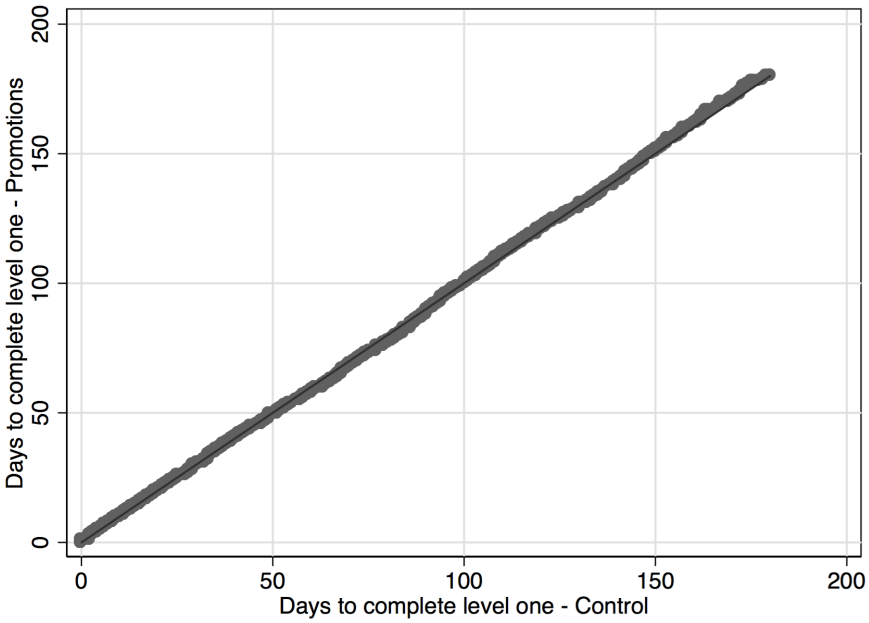
(a) In-game currency

(b) Sale of in-game currency

Notes: The top panel shows a screen shot of "gold bars," the in-game currency available for purchase in the popular game Candy Crush Saga. The user can buy the bars in various bundles. The lower panel shows a promotion run in the game for a bundle of in game currencies and other in-game features.

Days in app is balanced between treatment and control groups

Figure 2.11: Q-Q plot of days to complete level one in treated and control groups



Notes: Users in the treatment group have to complete level one before they are exposed to the promotional schedule. Since users are not exposed to any promotions before they reach level one, the set of users who reach level one in the treatment and control group remain balanced. Suppose a user i in the treated group reaches level one in τ_i days, so is exposed to the treatment for $d_i = 180 - \tau_i$ days. The figure presents a quantile-quantile plot of the distribution of d_i in the treatment (y-axis) and control (x-axis) groups to check balance. The distribution of d_i in the treatment and control groups is seen to be balanced.

Balance tests: Treatment vs. control groups

Table 2.3: Pre-treatment tests of balance for treatment and control groups

	Treatment: Promotions (T)		Control: No promotions (C)		
	$N = 134,061$		$N = 26,521$		
	Mean	SD	Mean	SD	p-value
Days to complete level one	9.88	(22.42)	9.84	(22.43)	0.771
Hours spent in app	8.47	(20.82)	8.52	(21.16)	0.756
Conversion	0.006	(0.078)	0.006	(0.074)	0.256
Purchases	0.008	(0.128)	0.007	(0.121)	0.192
Revenue	0.072	(1.38)	0.064	(1.35)	0.391
Connected to Facebook	0.295	(0.456)	0.300	(0.458)	0.131
Device memory	1,860.3	(869.5)	1,862.4	(869.3)	0.718
Device DPI (dots-per-inch)	303.6	(120.6)	303.3	(120.2)	0.701
Downloaded game in the US	0.32	(0.466)	0.32	(0.466)	0.92
Downloaded game in the UK	0.050	(0.218)	0.049	(0.217)	0.751
Downloaded game in Germany	0.153	(0.360)	0.154	(0.361)	0.728
Downloaded game in France	0.106	(0.307)	0.106	(0.308)	0.738

Notes: Pre-treatment indicators that are observed during users' play of level one of the game for control and treatment groups with p-values of t-test for difference.

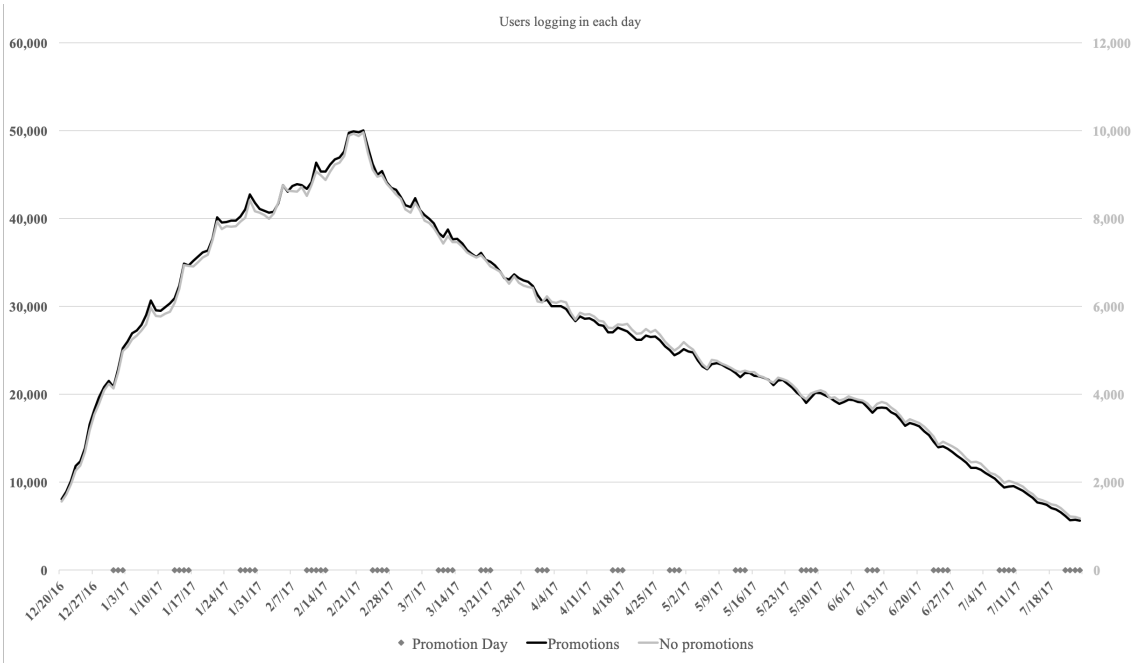
Balance tests: Three treatment sub-groups

Table 2.4: Pre-treatment tests of balance for treatment sub-groups

	Treatment: Immediate Promotions (T0)		Treatment: 25 Day Delay (T25)		Treatment: 50 Day Delay (T50)		
	$N = 54,883$		$N = 52,652$		$N = 26,526$		
	Mean	SD	Mean	SD	Mean	SD	p-value
Days to complete level one	9.83	(22.32)	9.90	(22.33)	9.95	(22.80)	0.773
Hours spent in app	8.55	(21.17)	8.49	(20.31)	8.26	(21.08)	0.172
Conversion	0.007	(0.081)	0.006	(0.077)	0.006	(0.074)	0.167
Purchases	0.009	(0.144)	0.008	(0.120)	0.007	(0.110)	0.034
Revenue	0.078	(1.51)	0.070	(1.33)	0.061	(1.16)	0.233
Connected to Facebook	0.296	(0.456)	0.295	(0.456)	0.295	(0.456)	0.973
Device memory	1,855.1	(871.0)	1,862.6	(866.5)	1,866.5	(872.6)	0.159
Device DPI (dots-per-inch)	302.5	(120.4)	304.4	(120.7)	304.2	(120.8)	0.025
Download in US	0.320	(0.466)	0.321	(0.467)	0.320	(0.467)	0.900
Download in UK	0.049	(0.215)	0.052	(0.221)	0.050	(0.217)	0.063
Download in Germany	0.156	(0.362)	0.151	(0.358)	0.151	(0.358)	0.090
Download in France	0.104	(0.306)	0.107	(0.309)	0.106	(0.308)	0.396

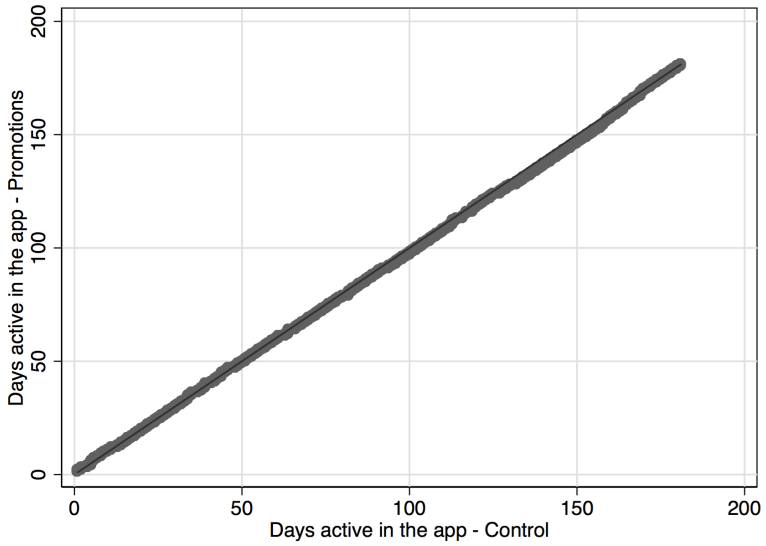
Notes: Pre-treatment indicators that are observed during users' play of level one of the game for all treatment sub-groups. p-values are derived from a three-way ANOVA test for differences.

Figure 2.12: Number of users logging into game by day split by group



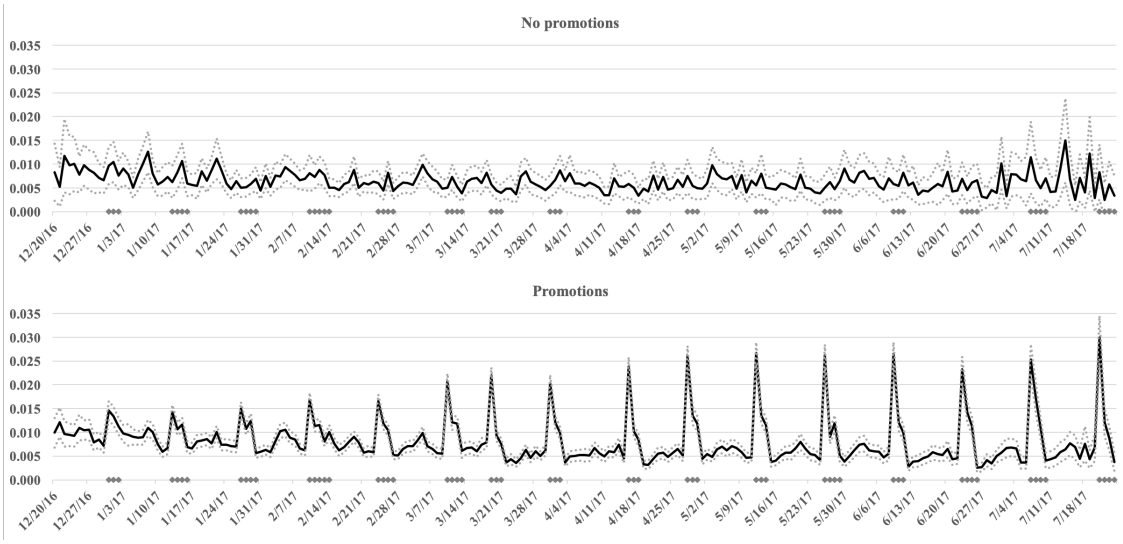
Notes: App use propensity is not affected by promotions: The number of users logging into the app remains virtually the same in treatment and control group (users in treatment group shown against the left, and users in the control group shown against the right y-axis).

Figure 2.13: Q-Q plot of days a user is active in the app split by group



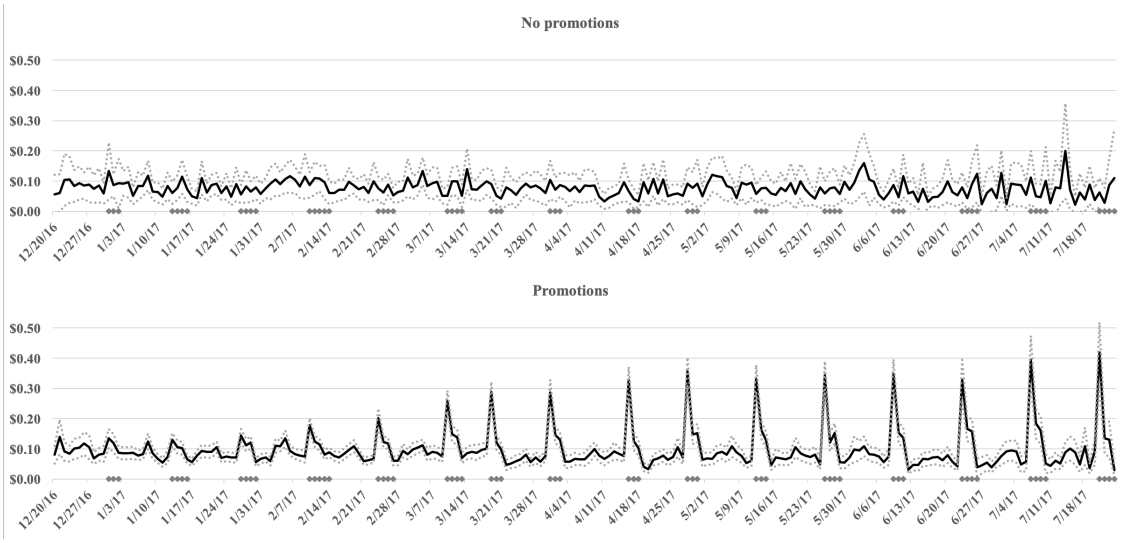
Notes: Days active is defined as the number of days (out of 180) that a user logged into the game. This Q-Q plot shows that the distribution of login propensity is the same across the treatment and control group.

Figure 2.14: Per-day purchases for treated and control groups



Notes: Purchases per user per day with 95% confidence band over calendar days, for the control (upper panel) and the treatment group (lower panel). Grey diamonds on the date axis indicate days when promotions were active.

Figure 2.15: Per-day revenues for treated and control groups



Notes: Revenue per user per day with 95% confidence band over calendar days, for the control (upper panel) and the treatment group (lower panel). Grey diamonds on the date axis indicate days when promotions were active.

Information about sales is widely discussed on online forums

Figure 2.16: Screen shots from a gaming forum

Notes: Screen shots from an online forum where information about sales in the video game is discussed before the experiment underlying this study was run. Identifying words are blacked out at the request of the data sponsor.

Chapter 3

Churn Prediction and Prevention in Freemium Apps: Can Free Premium Goods Change Churners' Mind?[1]

Julian Runge

Peng Gao (No affiliation)

Florent Garcin (Ecole Polytechnique Fédérale de Lausanne)

Boi Faltings (Ecole Polytechnique Fédérale de Lausanne)

Abstract

Predicting when users will leave an app creates a unique opportunity to increase their lifetime and revenue contribution. This paper focuses on predicting churn of high-value users and evaluates the ability of free premium goods to prevent churn in a field experiment. Offline evaluation compares the prediction performance of four common classification algorithms on datasets from two freemium gaming apps, each with millions of users. Furthermore, the authors implement a Hidden Markov Model to explicitly address temporal dynamics. Results indicate that a neural network achieves best prediction performance in terms of area under the receiver operating characteristic curve. The authors then conduct a field experiment with the neural network churn predictor in one of the apps to evaluate the impact of a free bundle of premium goods targeted to churning users: In a predictive treatment condition, users identified by the neural network predictor are targeted; in a heuristic treatment condition, all users who have not been active for 14 days are targeted; and in a control condition, no target offers are sent to users. Results from the experiment show that contacting users shortly before the predicted churn event substantially improves the effectiveness of communication with users. They further indicate that giving out free premium goods does not significantly impact the churn rate or future monetization of high-value users – suggesting that users can only be retained by remarkably changing their experience ahead of the churn event and that cross-linking or advertising may be more effective measures to deal with churning users in freemium apps.

[1] An earlier version of this paper was presented at and published in the proceedings of the Computational Intelligence and Games (CIG) conference 2014 which the present work references as Runge et al. (2014).

3.1 Introduction

The app economy has evolved to be a bustling marketplace where millions of firms vie for smartphone users' attention and wallets (Einav et al. 2014; Han et al. 2015; Arora et al. 2017). Most apps can be downloaded and used for free and premium upgrades are offered in in-app purchases – an extreme form of penetration pricing termed "freemium" (Niculescu and Wu 2014; Bapna et al. 2017; Gu et al. 2018). While many users only sample apps for minutes, hours or days, a few engage sustainably (Ghose et al. 2013; Hong and Pavlou 2014; Ross 2018). A subset of these sustainably engaged users goes on to spend substantial amounts of money on in-app purchases (Perez 2019; Sifa et al. 2018). Particularly games have been successful in adopting freemium pricing (often called "free-to-play" in that setting) and integrating with online social networks to reach and bind mobile users to their content offering (Alsén et al. 2016): Gaming apps do not only account for the majority of close to 200 billion app downloads in 2018 but also for three quarters of the revenue obtained on app stores (App Annie 2018; Sensortower 2019b).

Due to their disproportionately high contribution to value generation and profits, firms have major interest to retain sustainably engaged and monetizing users (Ross 2018; Sifa et al. 2018; Appel et al. 2019). The ability to predict when such a user will leave an app opens up an opportunity to adjust their experience to extend the lifetime in the app or to suggest another app to the user in hopes to "ignite" a new lifetime (Milošević et al. 2017). To achieve this, users can be incentivized to stick with an app, cross-linked to another app in the company's portfolio or cross-sold to other companies through advertising. In this paper, the authors design, implement and evaluate a churn prediction model for high-value users based on users' activity data in two very large gaming apps and test incentivization with free premium goods as a method to retain churning high-value users in one of the apps.

After defining the high-value user segment and the churn event for the two apps

under study, we formalize churn prediction as a binary classification problem. We then compare the prediction performance of four classification algorithms and explore the temporal dynamics of time series data using a Hidden Markov Model (HMM). In order to evaluate the impact that can be derived from targeting a free bundle of premium goods to churning users of an app, we design and implement a field experiment in one of the apps: In a predictive treatment condition users identified by the neural network predictor are targeted; in a heuristic treatment condition, serving as a naive comparison benchmark, all users who have not been active for 14 days are targeted; and in a control condition, no target offers are sent to users. Results indicate that contacting users shortly before the predicted churn event substantially improves the effectiveness of communication with users. They further show that giving out free premium goods does not significantly impact the churn rate or future monetization of app users – suggesting that users can only be retained by remarkably changing their experience ahead of the churn event and that cross-linking or advertising may be more effective measures to deal with churn of high-value users in freemium apps.

3.2 Conceptual Background

Two streams of literature are relevant to the present work: literature on the monetization of freemium goods and literature on consumer churn prediction and prevention. We will present a concise review of select papers from each in the following paragraphs.

Speaking to the first body of literature, Bapna and Umyarov (2015) and Bapna et al. (2017) present empirical studies that focus on social interaction as a central element of users' purchase intentions. Lambrecht and Misra (2016) study engagement as a key antecedent of users' purchasing of premium upgrades. Analytical accounts (Halbheer et al. 2014; Appel et al. 2019) emphasize how firms have to

trade off advertising and premium purchasing in light of user behavior dynamics. Runge et al. (2019) focus on the effect of regular promotions for the monetization of freemium offerings and find that promotions can lead to strong increases in primary demand. These studies highlight two factors in particular that make the monetization of freemium products challenging for firms: (1) Users tend to sample many free apps and only retain with a few, see also Ross (2018); (2) uncertainty about quality is high, making it difficult to entice consumers to open their wallets (e.g., Foubert and Gijsbrechts 2016; Appel et al. 2019). Our study speaks to this literature in studying churn prediction in a freemium setting and attempting to extend users' lifetime with an app by means of a free promotional bundle of premium goods as a retention incentive. The analysis focuses on high-value users who have produced substantial revenue for the firm as firm profitability in this setting hinges on the retention of such users due to their disproportionately high revenue contribution (Sifa et al. 2018). Premium goods should be well suited as an incentive for these users as they have previously revealed their preference for such goods by purchasing them.

Another longstanding stream of literature that is relevant to the present work concerns consumer churn and its prevention (Lemmens and Gupta 2020). The prediction of churn has been studied in various settings: e.g., see Burez and Van den Poel (2007) for the case of a television company, Ascarza et al. (2016) and Ascarza 2018 for the case of wireless communications and a special interest organization, Baumann et al. (2015) and Coussement et al. (2017) for the case of telecommunications, Coussement and Van den Poel (2008) for subscription services, Xie et al. (2009) for the case of banks, Hadiji et al. (2014); Rothenbuehler et al. (2015); Periáñez et al. (2016); Milošević et al. (2017); Banerjee et al. (2019) for the case of online games. The presented list is far from exhaustive – Ascarza et al. 2018 present a recent overview of and outlook for churn management research. The present paper addresses this literature in being the first to study churn prevention for engaged users

in the app economy (Han et al. 2015; Arora et al. 2017; Milošević et al. 2017) using an incentive that is uniquely suited to prevent churn in this setting. Methodologically, the authors focus on the application of machine learning techniques that have been reported as effective for churn prediction specifically (Burez and Van den Poel 2007; Coussement and Van den Poel 2008; Glady et al. 2009; Burez and Van den Poel 2009; Chen et al. 2012; Rothenbuehler et al. 2015; Milošević et al. 2017) and consumer behavior prediction more widely (West et al. 1997; Lemmens and Croux 2006; Briesch and Rajagopal 2010; Sifa et al. 2018). This aspect will receive further attention in Sections 3.3.4 and 3.4.4 later in the text.

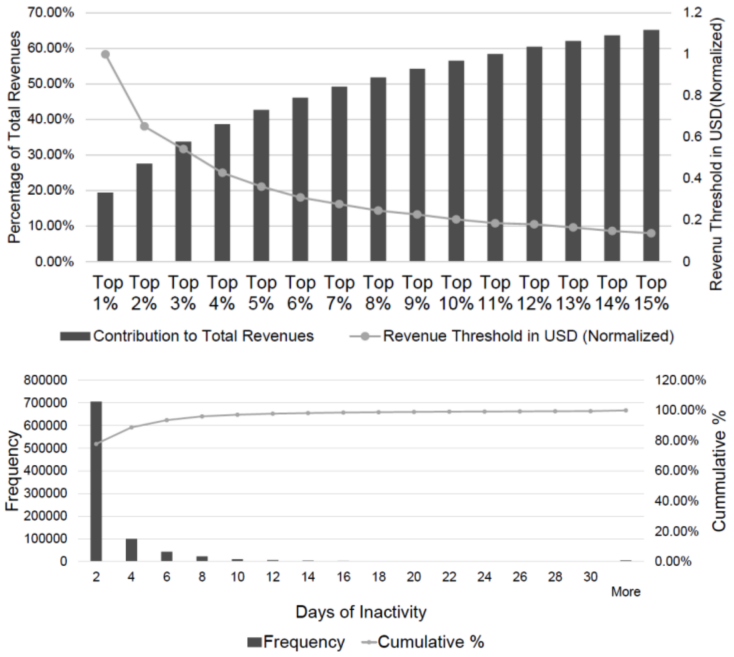
Summarizing, the questions guiding this study are: How can firms producing and marketing freemium apps identify high-value users at risk of churning? Can they retain such at-risk users by means of data-driven churn management? Are premium goods an effective incentive towards preventing churn? What can we infer as to good treatments to prevent consumer disengagement in this setting more generally?

3.3 Method

3.3.1 Empirical setting

The two gaming apps made available by the data sponsor for this study are published on the Apple App Store and the Facebook Appstore. The first app (called App 1 henceforth) is a puzzle game, like Tetris or Bejeweled, on mobile devices. The player has to clear gems of the same color as quickly as possible in a specified time in order to achieve a high score. Cleared gems are replaced from the top. There are also several power-ups or boosts present in the game that can make clearing the gems faster, so users can obtain a higher score and have a more exciting game experience. The second app (called App 2 henceforth) is a farming game which simulates the economics of running a highly personalized farm. The user plants

Figure 3.1: Revenue and activity distributions in the studied apps



Notes: The upper panel shows the contribution of the top x% of users to overall revenue (left y-axis) and the corresponding revenue threshold (right y-axis, normalized at the request of the data sponsor for confidentiality reasons). The lower panel shows what number (left y-axis) and cumulative share (right y-axis) of users is still active after x days of inactivity; the cumulative curve flattens post 14 days of inactivity, i.e., after 14 days of inactivity only 2% of users ever become active again.

and harvests crops to obtain coins and other in-game currency. As the users climb through various levels, they have the opportunity to extend and beautify their farm, and unlock new game features. Unlike App 1, App 2 is driven by missions that the players are asked to complete, requiring more regular interaction and engagement from users.

3.3.2 Definitions

The term "high-value users" is a rather vague term. In order to pin down a more precise definition for high-value users, we consider the contribution of top percentile paying users to the total revenue for the apps as shown in the top panel of Figure 3.1. The top 7% of paying users contribute around 50% of the total revenue. The minimum revenue threshold – the minimum revenue generated by a user in order to make it to the top percentile – starts to flatten out below the top 10% of

paying users. The top 10% seem to catch all users with an exceptionally high value which leads us to adopt the following definition:

> Definition 1. A high-value user in the app on day $t = 0$ (the observation time) is a user that ranks in the top 10% of all paying users sorted in decreasing order of revenue generated by each user between days $t = -90$ and $t = -1$.

Since the aim is to reach users before they churn from the app, the prediction should be targeting *active* high-value users. We therefore define active high-value users as follows:

> Definition 2. An active high-value player on day $t = 0$ is a high-value player who has played the game at least once between days $t = -14$ and $t = -1$.

The term "churn" is intended to capture that a user has permanently left the app. This decision may be conscious or not, driven by external or internal reasons. In practice, we need a threshold value for days of inactivity that we can use to clearly define the churn of a user. To this avail, we consider the distribution of days of inactivity between logins for high-value users in the apps. For example, if a high-value user played the game on day $t = 1$ and day $t = 3$, then again on day $t = 7$, this yields two samples of the days of inactivity: one is $3 - 1 - 1 = 1$ and the other is $7 - 3 - 1 = 3$. The bottom panel of Figure 3.1 shows the histogram of the distribution and cumulative distribution curve of days of inactivity. Less than 2% of high-value users stay away from the app for more than 14 days. Hence, 14 days of inactivity is a good indicator of churn. With this definition, 98% of the users defined as churners truly permanently disengage from the app.

> Definition 3. An active high-value user is said to be churning on day $t = 0$ if she starts a period of 14 consecutive days of inactivity on any of the days between day $t = 0$ and day $t = 6$.

3.3.3 Problem statement

We model the churn prediction problem as a binary classification task where the goal is to assign a label "churn" or "no churn" to each user. We train various classifiers on labelled data of previously observed player behavior up to a given day, and predict whether a player will churn or not within the week following that day. We use ROC-AUC (short: AUC), i.e., the area under the receiver operating characteristic (ROC) curve, for performance comparison because it allows to compare models across all possible classification thresholds. ROC curves are commonly depicted in a chart with the false positive rate (FPR) on the x- and the true positive rate (TPR) on the y-axis. Classifier performance can be compared for different combinations of TPR and FPR and hence for different threshold choices. As AUC is the area under the ROC curve, when it is one, the classifier performs perfectly and the ROC curve follows the left and top border of the chart. Regardless of the threshold, the TPR of the classification then is one and the FPR is zero. When the ROC curve is a diagonal from the lower left to the top right, AUC is 0.5 which reflects the case of random classification. Summarizing, our problem is to find the classifier that most correctly assigns the churn label to players across all possible classification thresholds and hence maximizes AUC.

Formally, with the above definitions, we can state the high-value user churn prediction problem as follows. Given all available historical tracking data of high-value users of an app, a training dataset D can be constructed in the following way: $D = \{(X_i, c_i)|i = 1, 2, \ldots, n\}$ where n is the number of users, each X_i in the dataset is a data instance vector consisting of historical activity data for an active high-value user up until a certain observation day, and c_i is a binary label indicating whether the high-value user is churning on the observation date. On this background, we can now formulate the churn prediction problem as a binary classification problem where we aim to find the binary classifier $f(x)$, trained by D, such that for an unlabeled dataset X_{new}:

$$f(x) = \arg\max\{E[AUC(f(X_{new}))]\} \tag{3.1}$$

3.3.4 Predictor selection

The churn classification problem that we formalized in Section 3.3.3 lends itself to the application of supervised learning algorithms whose ability to predict consumer behavior has been previously documented. Logistic regression (Logistic) is a workhorse model in the literature to predict binary choice (McFadden 1973). Similarly, neural networks (NNs) and decision trees (DTs) have been reported to work well for consumer behavior prediction in various settings (West et al. 1997; Lemmens and Croux 2006; Lessmann et al. 2015) and in this specific setting (Sifa et al. 2015; Hadiji et al. 2014; Milošević et al. 2017; Sifa et al. 2018). Finally, we include support vector machines (SVM) for their reported strength in churn classification problems (Coussement and Van den Poel 2008; Lessmann and Voß 2009; Chen et al. 2012). We apply all of these algorithms to approximate the unknown function $f(x)$ in equation (1) while maximizing predictions' AUC on holdout datasets through 10-fold cross-validation. In an attempt to better understand if latent temporal dynamics preceding the churn event can help predict it, we also implement a HMM to capture hidden user states (Burez and Van den Poel 2007; Netzer et al. 2008). We do not use these states for prediction directly (Rothenbuehler et al. 2015), but include them as features in the binary classification task described above (Burez and Van den Poel 2007).

3.4 Offline Evaluation: Predicting Churn of High-Value Users

3.4.1 Data preparation

For both apps, we extract the relevant historical tracking data of high-value users for two randomly chosen observation days, July 1st and August 1st of a recent year. We then construct two labeled datasets to build the churn prediction model.

Table 3.1: Offline datasets used for predictor evaluation

	Number of users	Number of churners	Churners over all users	Number of features in raw dataset	Final feature set
App 1	10,736	1,821	16.96%	516	Daily time series of rounds played, play accuracy, invites sent; days in game, last purchase, days since last purchase
App 2	7,709	352	4.57%	699	Daily time series of logins, level; in-game currency 1 balance, in-game currency 2 balance

Notes: App 2 has a substantially lower churn rate as users retain longer on average; for both apps, the final datasets after feature selection contain usage- and purchase-related features.

Table 3.1 shows a summary of the dataset. There are three main categories of data: first, in-app activity tracking data such as a time series of logins per day or a time series of play accuracy; second, revenue-related tracking data, such as a time series of revenues generated or in-game currency balances held by users; third, user profile data, such as how long the user has been using the app and which country the user is from. We process the data to alleviate high positive skewness of the datasets by applying a Box-Cox transformation. Since the selected prediction algorithms usually expect standardized input data, we further center and scale the data during the data preparation phase.

The prepared datasets include more than several hundred attributes, and not all of them are likely to be informative for making predictions. To identify the set of attributes to be used for prediction, we perform a series of feature selection procedures. We use logistic regression with 10-fold cross validation to estimate the AUC performance of different feature sets. We experiment with the length of time series, eliminate time series that are highly correlated with others, and apply forward feature selection. The last column of Table 3.1 summarizes the feature sets retained the final models. Empirical experiments on offline data further suggest that using the last 14 days of historical data prior to the churn event yields the best prediction performance in terms of AUC.

Table 3.2: Mean AUC achieved by prediction algorithms

	Neural network	Logistic regression	Decision tree	Support vector machine
App 1	0.815	0.814	0.732	0.850
App 2	0.930	0.924	0.850	0.903

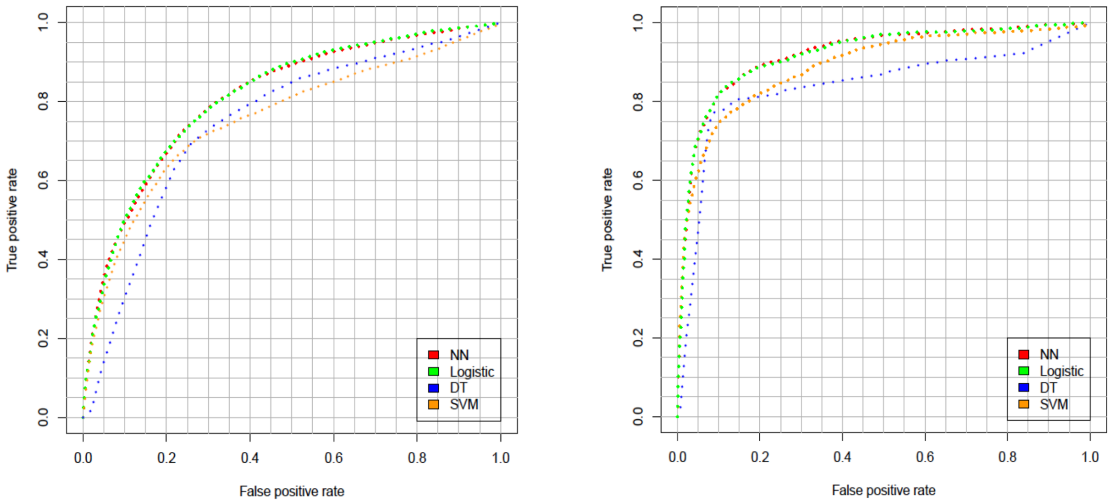
Notes: All algorithms achieve substantially higher prediction performance for App 2 than for App 1; a neural network predictor achieves best performance on both apps' datasets.

3.4.2 Offline evaluation of prediction algorithms

Prior to implementing an online churn prediction system, we compare the prediction performance of our selected algorithms. For the SVM, we use the radial basis function kernel and apply a parameter grid search to tune the hyper-parameters with 10-fold cross-validation. We experiment with 100 combinations of C (ranging from 0 to 5) and gamma (ranging from 0.2 to 5) with quadratic step size. The NN has a simple one-hidden layer network topology. The number of hidden nodes is set to be equal to the sum of the number of attributes and classes divided by two plus one. Also for the NN, we use a parameter grid search with 10-fold cross-validation and try 100 combinations of learning rate and momentum (both from 0.5^1 to 0.5^{10}).

Table 3.2 reports average AUC performance of different algorithms over the two datasets. Figure 3.2 shows the ROC curves for the four prediction algorithms on the two apps' datasets. Results are consistent over the two datasets. The performance of NN and logistic regression tracks closely, but the NN is slightly better based on mean AUC. The performance of SVM and DT falls behind logistic regression and NN. SVM performs better for low FPRs and DT performs better for FPRs higher than 25%. In application of a churn prediction system, a low FPR is more important because accidentally treating non-churning users can be more costly than missing some of the churners. This is so because non-churners have a higher expected future revenue contribution than churning users – whose future value is by definition zero. Hence, the SVM would be preferred to the DT. Also in terms of AUC the SVM outperforms the decision tree.

Figure 3.2: ROC curves of the four selected prediction algorithms for both apps



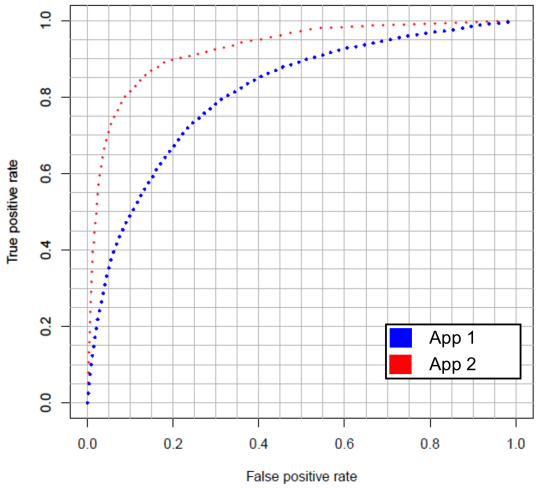
Notes: App 1 on the left, App 2 on the right; neural network and logistic regression track closely and outperform both the SVM and the decision tree predictor.

3.4.3 Prediction performance across the two apps

The previous section established that the ranking of prediction algorithms in terms of AUC is the same across both apps – the NN provides best performance in terms of mean AUC. To understand how prediction performance compares between the two apps, Figure 3.3 shows a comparison of the ROC curves for a NN on App 1's versus App 2's dataset. The same prediction modeling technique performs much better for App 2 than for App 1. More specifically, if we fix the FPR – that is the percentage of actually non-churning users we include in the predicted list of churns – at 5%, we achieve a TPR higher than 70% for App 2. Hence, we reach more than 70% of truly churning users, while for App 1, we only reach 35% of truly churning users for the same FPR.

An intuitive explanation for this difference is that the nature of the apps differs as discussed in Section 3.3.1. App 2, though being a freemium game as App 1, requires more constant interaction from users and is characterized by higher and more constant engagement. It offers many additional experiences on top of the core farming mechanic. Among these are crafting and selling products, lotteries, an underwater garden and deep social features like visiting friends' farms. App 1

Figure 3.3: ROC curve of neural network predictor for both apps



Notes: A neural network predictor performs best for both apps as shown in Figure 3.2; its prediction performance is much stronger for App 2 than for App 1.

on the other hand fully focuses on timed rounds of the same core mechanic. It does not require a high level of commitment from users and allows for more casual interactions. Logins per user per day are substantially lower in App 1 than in App 2 and times between sessions for one user can span several days.

3.4.4 Combining neural network and HMM

Though the modeling techniques discussed thusfar already deliver good prediction performance, one common issue with all the techniques considered is that they do not take the temporal dynamics of the time series attributes into consideration explicitly. If we switched the order of the data points in the time series, the resulting prediction would not be affected since (time-wise) ordering of attributes is not accounted for by these algorithms. There are high quality historical tracking data dating back months and years available in the data sponsor's databases. In order to better leverage the information potentially present in these data, we turn our focus to HMMs: We include the results obtained from a HMM in the neural network to further improve the prediction performance. We focus on App 2's dataset to further develop this line of thinking as predictors achieve better baseline performance on it compared to App 1's dataset.

The data under study are all instances of the logins per day time series for App 2 after the data cleaning step but without data transformation. This is because the data transformation alters the data in a way that makes them unusable for fitting a HMM. The training data for the HMM can be denoted as a vector

$$L = [L_1, L_2, \ldots, L_n]^T \tag{3.2}$$

where each $L_i = [L_i(-60), L_i(-59), \ldots, L_i(-1)]$ is the time series of logins per day for a user between day $t = -60$ and $t = -1$ and n is the number of instances in the dataset.

We make the following assumptions regarding the model:

- The instances of the logins time series are mutually independent which is a valid assumption since each instance is an observation of a certain different user.
- All instances of the logins time series are generated from one single underlying hidden Markov process. We hence assume that the HMM portrays an average user's stochastic behavior.
- The emission distribution of the HMM follows a Poisson distribution. Each value in the logins time series is a non-negative integer that records the number of login events.

Essentially the model setup reflects that the actual logins of a user on a certain date depend on the states of all users on that date, and that the state process – which is hidden and unobservable – is a Markov chain process. The actual observed values of logins follow a Poisson distribution, where the mean of the Poisson distribution depends on the state of a user on each date.

With the HMM, we leverage more historical data and take temporal dynamics explicitly into consideration. However, the HMM setup is hard to reconcile with our definition of churn and cannot be used directly for making predictions. In order to

make use of the HMM that we devise, we use it to extract new features to be added to the NN predictor. The idea is that this will enhance the prediction performance. We follow the approach of Burez and Van den Poel (2007) who incorporate features extracted from a Markov chain model into a NN to improve prediction performance. Through the HMM we calculate the following probabilities for each instance i in D as new features to add to the neural network:

$$p_i = [p_0, p_1, \ldots, p_{13}] \tag{3.3}$$

where $p_k = Prob(L_i(k) = 0 | L_i = l_i)$. Essentially, p_i is a vector with element p_k being the probability of user i not using the app on date $t = k$ given the observed sequence of L up to $t = -1$.

Under the setup of the HMM, the probability p_k can be easily calculated using the following equation:

$$p_k = (\alpha \Gamma^{k+1} P(0) 1')/(\alpha 1') \tag{3.4}$$

where α denotes the forward probabilities of the HMM at $t = -1$, Γ is the transition matrix of the HMM, and $P(0)$ is a diagonal matrix where the m-th diagonal element is the probability of observing a 0, given the hidden state is m. Detailed mathematic proof of the above equation can be found in Zucchini et al. (2017).

We then add the new features, p_i, into App 2's dataset after transformation and apply the neural network modeling on top of the new feature set. The AUC value of the model with the new feature set is 0.923, which degrades the prediction performance compared to the 0.930 achieved with the feature set not including the users' hidden state. Only using the HMM features for prediction yields a mean AUC across 10 cross-validation runs of 0.915. Inclusion of the HMM features hence does not appear to add incremental prediction performance and we do not include HMM results in our final prediction model for the online experiment described in the next section.

3.5 Online Experiment: Targeting Free Premium Goods to Churning Users

Now that we identified a well performing churn prediction model, we can apply it online to identify churning users and try to prevent their disengagement from the app. As prediction performance is substantially better for App 2 compared to App 1, the data sponsor encouraged us to use App 2 for an online application of the predictive churn system. Prior to describing the experimental design, we introduce the treatment to prevent churn in more detail.

3.5.1 Experiment design

Freemium apps monetize through the sale of premium upgrades through in-app purchases. The goods sold in such purchases generally complement and enhance the free product experience, e.g., by providing game enhancements, extended functionality or more exciting experiences (Hamari and Keronen 2017; Hamari et al. 2020). Much of the user experience in freemium apps is structured around such purchases and intends to create need and desire on the user's part to make a purchase (Lehdonvirta 2009; Hamari et al. 2020). It hence seems that the premium goods sold should make for a good incentive to retain users. This is expected to be particularly true when targeting high-value users – as in this study – because these users have previously revealed their preference for such goods by purchasing them in large amounts. In collaboration with the data sponsor, the authors hence devised a free bundle of premium in-game goods, worth more than $10 and hence more than three times the average daily spend of high-value users, as an incentive to retain high-value churners in App 2.

The online experiment has three main goals: (1) Apply the churn prediction model identified in offline evaluation in the field, (2) assess the effectiveness of the free premium goods incentive, (3) assess the value of a machine learning-based ex-

pert system over a simple heuristic implementation that managers could implement without predictive analytics. To achieve all these goals, we design an experiment with two treatment conditions and a control condition. Users in condition A are treated with the incentive using a simple heuristic based on definitions 2 and 3 in Section 3.3.2 that managers could implement without data scientists' help: All high-value users who have been inactive for 14 days are contacted with a Facebook application notification and an e-mail offering the free bundle of premium goods. Users in condition B are treated using a machine learning-based expert system: All high-value users who are predicted by our neural network predictor to churn within the next week (see definition 3 in Section 3.3.2) receive the free bundle of premium in-game goods. Condition C finally serves as the control condition: High-value users in this condition are not treated.

We use the following metrics to evaluate the impact of our treatments on high-value users:

> Definition 4. Churn Rate (CR) CR = 1 – (# active high-value users) / (# high-value users)
>
> Definition 5. Daily Revenues (DR) DR = total revenues from users in a group during a day
>
> Definition 6. Email Click Through Rate (CTR) CTR = (# gifts claimed by email) / (# emails delivered)
>
> Definition 7. Facebook Click To Impression Rate (CTI) CTI = (# gifts claimed by notification) / (# notifications seen by users)

Towards an evaluation of our treatment conditions' impact on these metrics, we randomize high-value users in App 2 into the different conditions. 40% of users go to condition A and B respectively, the remaining 20% receive no tratment and establish the control condition. The experiment comprises all users meeting the definitions put forth in Section 3.3.2 between mid-January to mid-February of a recent year.

Table 3.3: High-value user disengagement behavior throughout the experiment

		Condition A Heuristic targeting	Condition B Predictive targeting	Condition C No targeting
Start of experiment	Number of high-value users	1,717	1,607	789
	Active	1,583	1,472	744
	Churn rate	7.8%	8.4%	6.8%
End of experiment	Number of high-value users	2,034	1,896	933
	Active	1,809	1,680	837
	Churn rate	11.1%	11.4%	10.3%
Differential	Number of high-value users	317	289	135
	Active	226	208	93
	Churn rate	3.3%	3.0%	3.5%

Notes: The number of high-value users is increasing in each condition as is their churn rate. In line with expectations, the churn rate increases least in the predictive targeting condition and most in the condition where no retention incentive is targeted.

Table 3.3 shows how many users are in each treatment condition at the beginning and at the end of the experiment.

3.5.2 Results: Churn and monetization

Table 3.3 shows the number of treated users and churn rate of high-value users in different experimental conditions, observed at the beginning and the end of the experiment. As can be seen, churn rates of high-value users increased throughout the experiment duration for all three experimental conditions. This result is expected as churn dynamics change, e.g., over the lifecycle of an app, due to seasonality, changing user preferences and depending on the competitive landscape. Rather than absolute churn rate at the beginning or the end of the experiment, we focus on the differential between end and start for each condition as this number directly reflects the effectiveness of the different treatments in lowering churn: A more effective treatment should lead to a lower increase or a stronger decrease in churn over time. The last row of Table 3.3 shows that churn increased least in the predictive treatment condition B (+3.0%), followed by the heuristic treatment condition A

Table 3.4: Statistical significance of differences between conditions

Metric	Churn rate (differential)		Norm. daily revenue (t-test)		E-mail CTR	Notification CTI
Comparison	A vs. C	B vs. C	A vs. C	B vs. C	A vs. B	A vs. B
Test statistic	0.2636	0.4282	1.0067	0.7767	4.5569	24.1018
p-value	0.6077	0.5145	0.2909	0.4407	0.0328	0.0000

Notes: Pairwise comparison of outcomes in treatment conditions: Churn and revenue are not significantly lifted under either churn treatment compared to the holdout condition; communication effectiveness, as measured by e-mail CTR and notification CTI, is significantly lifted in the predictive compared to the heuristic treatment condition.

(+3.3%) and the holdout condition C directionally saw the strongest increase in high-value user churn (+3.5%). This ranking is in line with the expectation that a churn prediction expert system is best at administering an "anti-churn" incentive. Based on the assumption that user churn is a Bernoulli random variable we apply a Chi-square test to assess whether differences between conditions are statistically significant. As shown in column two and three in Table 3.4, differences are not statistically significant. Results hence only provide indicative directional evidence for the effectiveness of our predictive churn prevention treatment.

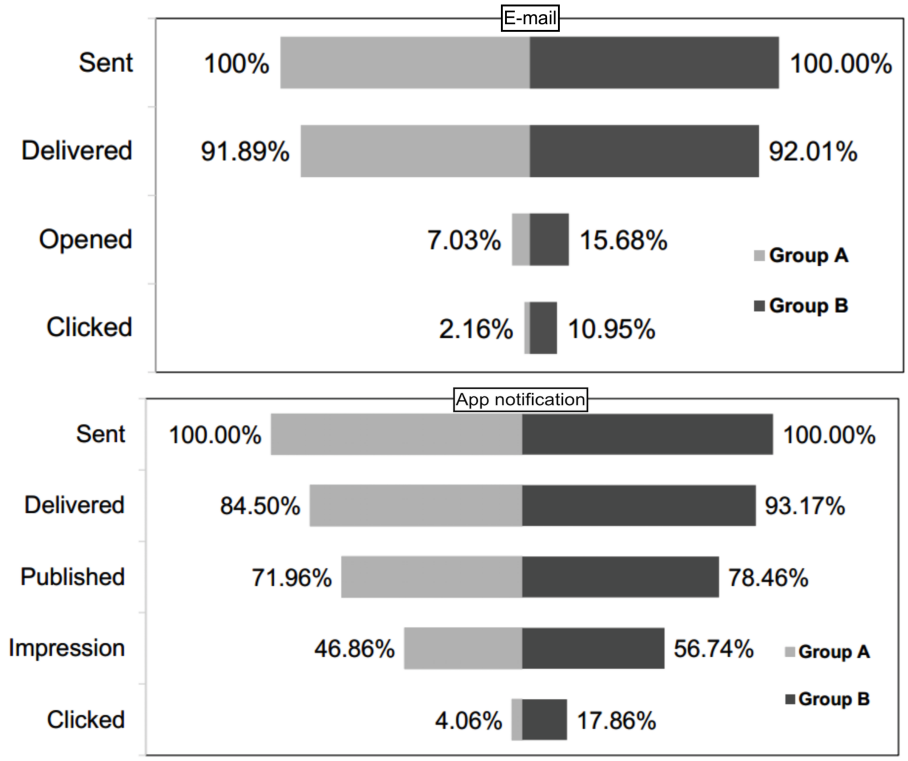
Historically, for the app under study, the normalized daily revenue follows a Gaussian distribution and has close to equal variance. Based on these characteristics, we perform a two-sample t-test to assess whether there is a significant difference in daily revenue between different experimental groups. The null hypotheses are that the means of the normalized daily revenues are equal between condition A and C and between condition B and C. Results are shown in Table 3.4. Since the p-values are again both substantially greater than 0.05, we are not able to reject the null hypothesis of equal means. Hence, there is no statistically significant evidence that daily revenue is affected by the different churn management policies.

3.5.3 Results: Communication effectiveness

While effects on churn and monetization are small or not detectable, Figure 3.4 shows that communication effectiveness is much higher in the predictive than in the heuristic treatment group (there was no communication to users in the control condition). In the heuristic condition (Group A), the CTR of the e-mail campaign is 2.4%. In the treatment condition using the churn prediction system (Group B), the same value is 11.9%. Table 3.4 shows Chi-square test results for comparing the e-mail CTR between Group A and Group B. The p-value of 0.0328 (< 0.05) indicates that with the prediction model, we are improving the effectiveness of e-mail marketing and seizing the opportunity to reach high-value users while they are still interested in the app. To further compare the funnel of the e-mail campaign of Group A and Group B, Figure 3.4 illustrates the difference between the groups. The top of the funnel notes the overall e-mails that were sent out. On the next step, in the delivery phase, the e-mail funnel in the top panel of Figure 3.4 loses about 10% of users for both groups since only about 90% of recorded e-mail addresses are valid. For the next step, 'Opened,' a clear difference in the opening rate becomes apparent. Finally, the predictive treatment group exhibits more than five times the CTR compared to the heuristic treatment group, and four times more users claim the gift links under the predictive churn management policy.

Similarly, the Facebook notifications CTI is 8.7% for the heuristic Group A while it is 31.5% for the predictive Group B. The p-value of a Chi-square test on the two groups is approximately 0 and therefore the predictive churn management system performs statistically significantly better in reaching users. The bottom panel of Figure 3.4 illustrates a comparison of the Facebook notifications funnel between Group A and Group B. The first step summarizes all users to whom the system sent out notifications via Facebook. For the next step, there is a difference in the delivery rate since we are not able to deliver notifications to users who have uninstalled the app. On the next level, 'Published,' further users are lost; the reason for this is

Figure 3.4: Communication effectiveness in the heuristic (A) versus the predictive (B) treatment condition



Notes: The firm's communication efforts reach users much more effectively in the predictive than in the heuristic treatment condition. E-mail funnel in the top, Facebook application notification funnel in the bottom panel. Group A received the heuristic, Group B the predictive targeting approach.

that if the user has blocked the app's notifications, Facebook will not publish any notifications to the user. For Group A and Group B, the percentage loss at this step is approximately equal. At the impression level, for Group A, only $46.9/72.0 = 65.1\%$ of the published notifications are seen by users while for the predictive Group B, $56.7/78.5 = 72.3\%$ published notifications are seen by users. In the final step, we witness a sizable improvement in the click rate of notifications under predictive churn management. Notifications are more than four times more likely to be claimed in the predictive churn management policy Group B compared to the heuristic treatment in Group A.

3.6 Discussion

3.6.1 Merits of the predictive churn management system

Our field experiment presents directional evidence that proactive outreach to at-risk users with a meaningful incentive can lower churn rates of these users: The churn rate in the predictive treatment group increases 14.3% less compared to the increase in the control condition (+3.0% versus +3.5%, see Table 3.3 on page 62) from beginning to end of the experiment.[2] This difference in increase is however not statistically significant ($p = 0.5145$, see Table 3.4 on page 63). There are further no detectable effects on monetization of app users from either heuristic or predictive churn management vis-a-vis the control condition.

While effects on churn and monetization are limited, the predictive churn management strongly (by factor four to five) increases effectiveness of communication with users. As shown in Figure 3.4 and Table 3.4, effects are both substantively large and statistically significant. A reasonable criticism of these results is that the good CTR achieved for the predictive group might be driven by false positives, i.e., users who are not about to churn but are predicted to do so. When training and testing our algorithm, we consistently achieved a precision of better than 40% for repeated and out-of-sample testing. Assuming the same precision for the prediction used in the field experiment, up to 60% of contacted high-value users may have been actual non-churners. In previous e-mail gift campaigns with high-value users for the same app, the firm observed CTRs of around 10%. Assuming a similar CTR for the false positives among the predicted churners, the true positives – i.e., users who are actually about to churn – have approximately the same CTR, since the overall CTR is above 10%. Hence, contacting high-value users shortly before they churn appears to be as efficient as contacting them earlier in their lifetime when they are still far

[2]Note that churn rates across all treatment conditions increased throughout the experiment duration of one month due to extraneous factors such as changing consumer preferences, competitive environment and seasonality.

away from churning. It further is much more efficient than contacting users after their churn event as implemented in the heuristic treatment condition – where we observe a CTR of around 2% (see Figure 3.4).

The value of the presented churn management system therefore lies in enabling the firm to contact users just before the end of their lifetime in the app when they are still fully responsive to the firm's communication efforts. While this is a valuable use case, results further indicate that a more meaningful incentive may be necessary to entice high-value users to change their behavior and "reignite" their interest in the app – or a different treatment altogether. We will address this line of enquiry in the next section.

3.6.2 How to treat churning users in freemium apps?

From a conceptual perspective, the incentive chosen by the authors – free premium goods worth more than three time the daily average spending of the targeted users – seems well suited: Freemium apps tend to be organized around the creation of interest in premium upgrades (Sifa et al. 2015; Levitt et al. 2016; Hamari and Keronen 2017; Runge et al. 2019) and, in financially successful apps, such premium upgrades usually present a noticeable improvement to the user experience (Lehdonvirta 2009; Perez 2019). Further all targeted users have made purchases of the premium goods in the past, meaning these goods provide(d) utility to them.

An argument explaining why the incentive is not more effective in retaining users lies in its explicit and extrinsic nature: Users have to consciously claim the free promotional bundle of premium goods. Doing so requires cognitive attention and the exertion of effort – open the e-mail/notification, process information, click to claim the bundle. Extrinsic incentives can crowd out intrinsic motivation (Deci et al. 1999). To the extent that users' engagement with the app derives from intrinsiv motivations, an increase in intrinsic usage incentives may be a more effective way to prevent churn. The firm could, e.g., automatically add the free premium goods to

users' accounts and notify them of the gift. Generally, incentives that are automatic and intrinsic may be more effective in this setting as users are intrinsically motivated to use an app (Deci et al. 1999; Lewis et al. 2012; Eyal 2014). Along these lines, other promising incentives might be to enhance the app experience without notifying the user well ahead of the churn event until the user is fully "hooked" and motivated to use the app again (Deci et al. 1999; Eyal 2014).

Another explanation for the limited effectiveness of the chosen incentive to retain users is that users just do not obtain utility from the premium goods any longer and that their preferences have irreversebly changed. Following this line of thinking, more promising approaches to deal with churning users are to suggest another app to them. Most firms in the app economy own and market more than one app (Han et al. 2015). The firm can hence crosslink a user to another app in its portfolio when the user is predicted to disengage from the focal app. Crosslinking users who are at the end of their lifetime and would leave the focal app anyway comes at virtually no cost. Finally, if the firm has no (suited) further apps in its portfolio, it can sell churning users to other firms via in-app advertising as a last resort (Halbheer et al. 2014; Appel et al. 2019). Future research can fruitfully address these questions and refine churn treatment policies through further analysis and experimentation.

3.6.3 Limitations and future research

The authors wish to end by naming relevant limitations and avenues for future research. A limiting element of the present study's offline evaluation is that it considers mean AUC without deriving confidence intervals. While confidence intervals will not change which predictor is selected for the online experiment, they would be insightful in establishing if differences between predictors are statistically significant or not. A further limitation is that features are selected based on AUC achieved with logistic regression. To the extent that other algorithms, e.g., DT, NN or SVM, might have benefitted more from certain attributes than logistic regression,

this could lower the performance achieved with these other predictors (Crone et al. 2006; Coussement et al. 2017). The main result, i.e., that a neural network is best able to predict churn in terms of AUC, is however robust to this limitation as more tailored feature selection would only improve (and not lower) the NN predictor's performance. Future research could address these limitations.

Further avenues for future research are to extend the present study's findings to more apps and other freemium offerings (e.g., news websites) and to refine churn treatment policies through further analysis and experimentation. When doing so, researchers might benefit from predicting the effectiveness of different incentives and treatments directly (Ascarza et al. 2016; Ascarza 2018). Different treatments may be most effective for different sub-segments of churning users and evaluating their effectiveness directly could identify more optimal churn management policies (Ascarza et al. 2018). Methodologically, reinforcement learning-based approaches (Rana and Oliveira 2015; Aboussalah and Lee 2020), e.g., contextual bandits (Li et al. 2010; Bietti et al. 2018), could assist with this identification by automatically learning targeted treatment assignment policies while maximizing a pre-specified reward. Such a reward ideally would proxy future revenue or future total logins of users directly.

Chapter 4

Early Detection of Paying Users in Freemium Apps: An Application of Deep Learning and Synthetic Oversampling[1]

Julian Runge

Rafet Sifa (Fraunhofer IAIS)

Christian Bauckhage (Fraunhofer IAIS & University of Bonn)

Daniel Klapper (Humboldt University Berlin)

Abstract

In freemium marketplaces, a small share of paying users drives firms' revenue. Marketers wish to identify such premium users early on to ensure their experience is rewarding and to direct advertising spend towards channels that yield promising users. This identification is complicated by the relative rarity of premium users and the fact that future "free" and paying users can display similar app use behavior, creating various non-linear associations in consumer behavior. The present study shows that synthetic oversampling of the minority class of premium users can help various learners in identifying such users. Further, neural networks should be particularly well suited for this prediction as they can fit most classes of linear and non-linear functions to arbitrary precision, especially when their architecture is "deep," i.e., has several hidden layers. Indeed, a neural network with four hidden layers – trained on oversampled data – surfaces as the best detector of future paying customers from app users' digital footprint. The presented methodology promises to have valuable applications for the identification of high-type users, particularly when samples are large, input data diverse, choice paths varied and imbalance in behavioral outcomes high.

[1] An earlier version of this paper, focusing on the implementation and benefits of synthetic oversampling, was presented at and published in the proceedings of the Hawaii International Conference on System Sciences (HICSS) 2018. It is referenced here as Sifa et al. (2018).

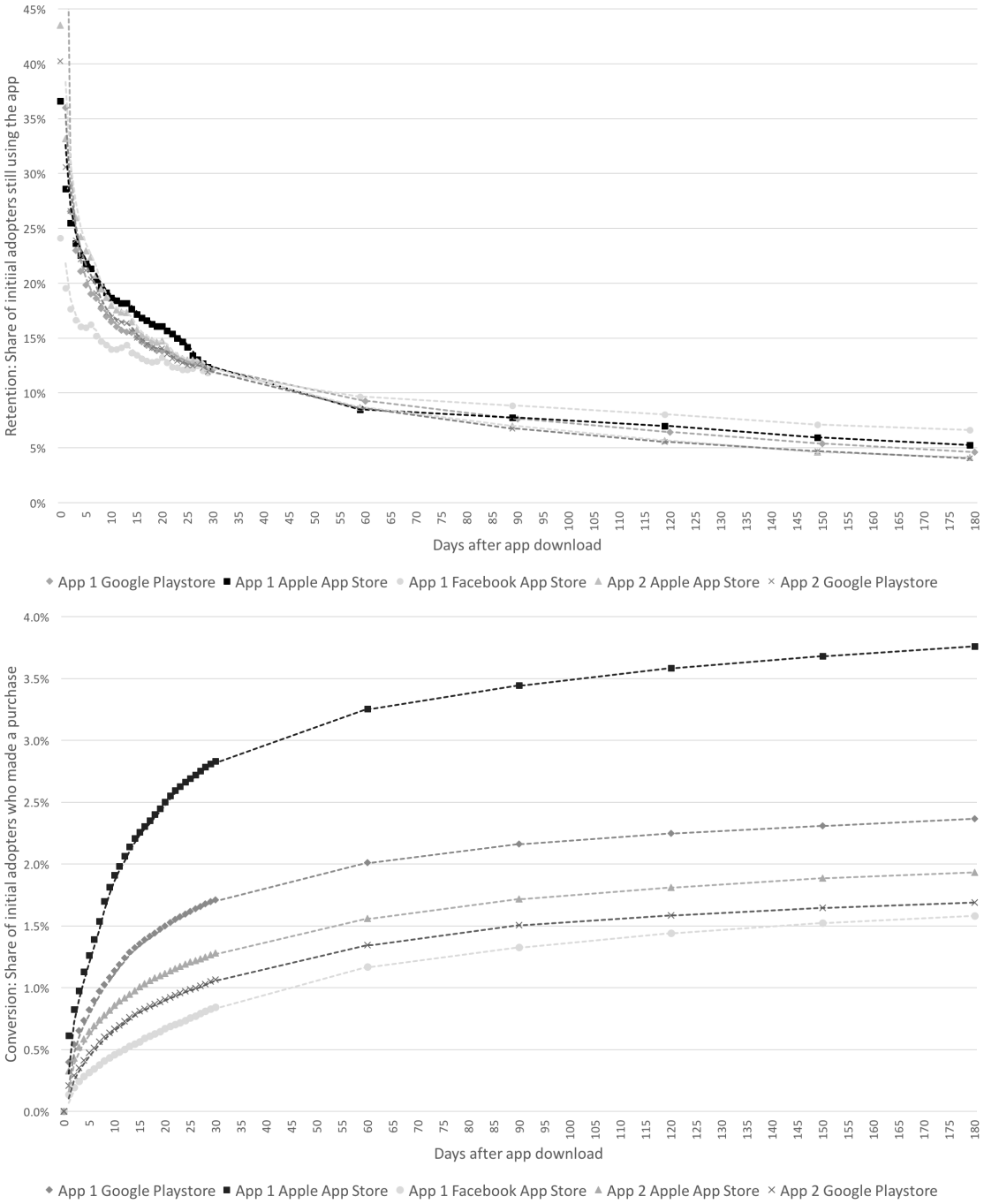
4.1 Introduction

Freemium has become a widespread pricing model for digital consumer goods: A basic version of the product or service can be used for free and premium upgrades are available against payment of a fee (Pauwels and Weiss 2008; Gu et al. 2018; Shi et al. 2019). It comes in two key flavors: Contractual and non-contractual. The former are subscription-based and used by, e.g., Dropbox, Spotify and many providers of digital news such as the New York Times or the Economist. Here, consumers upgrade a single time to a premium plan and then pay in monthly installments (Lee et al. 2017). Some firms have moved to hybrid models that extend contractual with non-contractual premium upgrade options; Tinder for example offers a 'Gold' subscription and numerous non-contractual upgrades such as additional 'Likes' and other virtual goods (Lehdonvirta 2009).

Non-contractual freemium pricing models are common in interactive online environments where users sample many products and only stick with a few. The app economy – a $143 billion industry with 12 million competing app developers according to 2016 figures (Arora et al. 2017) – is a good example, with the largest part of adopters disengaging within a week (see Figure 4.1).[2] In such environments, premium upgrades are usually offered in in-app purchases that tend to be repeated manifold by engaged users who drive the largest share of firms' revenue (Ghose and Han 2014) and finance the free provision of the product or service for the rest of users (Bapna et al. 2017). Managers strive to provide the best possible experience to these users (Berger and Nasr 1998; Malthouse and Blattberg 2005). In light of the short user lifecycles in freemium settings (Figure 4.1), early experiences can have crucial impact on consumers' adoption decision and marketers wish to identify paying prospects early on to take suited action to foster their retention and monetization.

[2]Note that our use of the term "retention" refers to app use after app download regardless if the user made a purchase or not. This use is in line with nascent literature on the app economy (Appel et al. 2019) and routed in the managerial use of the term (Ross 2018).

Figure 4.1: Retention and conversion in freemium mobile apps



Notes: Average share of a user cohort retained (i.e., still actively using the app either paid or for free, top panel) and average share of a user cohort that bought a premium upgrade (conversion, bottom panel) over days after initial app download for select apps of the data sponsor. Retention curves are representative of what has been reported for apps (Schonfeld 2009), many users only sample apps for a few hours or days. Conversion curves display stronger between-app differences than retention curves. The analysis in this paper is conducted on detailed data from the app with the median conversion profile.

To assist managers with this challenge, the present study zeroes in on the prediction of consumers' future premium demand from early behavioral traces as recorded shortly after initial app download. Interactive digital products such as mobile apps allow firms to log detailed records of consumer behavior in varied data formats similar to the clickstream data produced on websites (Bucklin et al. 2002; Bucklin and Sismeiro 2009). While this provides a rich dataset, prediction is complicated by the fact that "free" users can display behavioral patterns that are highly similar to the ones of (future) premium customers (Moe and Fader 2004; Shampanier et al. 2007). Figure 4.2 on page 76 shows behavioral traces during the first week after app download for future free and premium users. While future premium users register more sessions, rounds and days played and spend more time in the app on average, distributions of app use behavior between future free and premium users overlap substantially, suggesting the presence of decision rules of the form "I will never spend anything on this app, regardless how much I use it." Shampanier et al. (2007) document the difficulty to remove consumers from a zero-price point which manifests itself here in a "zero-lock-in" of users who appear behaviorally (in terms of app use, see Figure 4.2) similar to premium customers.

When consumer behavior shows such non-linear associations that can, e.g., also be caused by non-compensatory decision rules (Einhorn 1970), neural networks (NNs) have been found to perform particularly well for choice prediction (West et al. 1997). Due to their universal approximation property, they can fit most classes of linear and non-linear continuous functions to arbitrary precision (Hornik 1991), especially when their architecture is "deep," i.e., has several hidden layers (LeCun et al. 2015). The present study considers NNs with one to four hidden layers – and hence comprising both deep and non-deep topologies – to predict future premium demand from early behavioral traces of mobile app users and compares their performance to linear and random forest (RF) predictors. In doing so, the study is the first to present a detailed application of deep learning to consumer behavior prediction

and intends to help open marketing practice and research to this class of learning algorithms that has achieved breakthrough success in other domains (LeCun et al. 2015). The presented methodology promises to have valuable applications beyond freemium marketplaces, in the identification of high-type users on diverse behavioral outcomes from consumers' digital behavioral traces. It is likely particularly useful when samples are large, input data formats diverse, choice paths varied and imbalance in behavioral outcomes high.

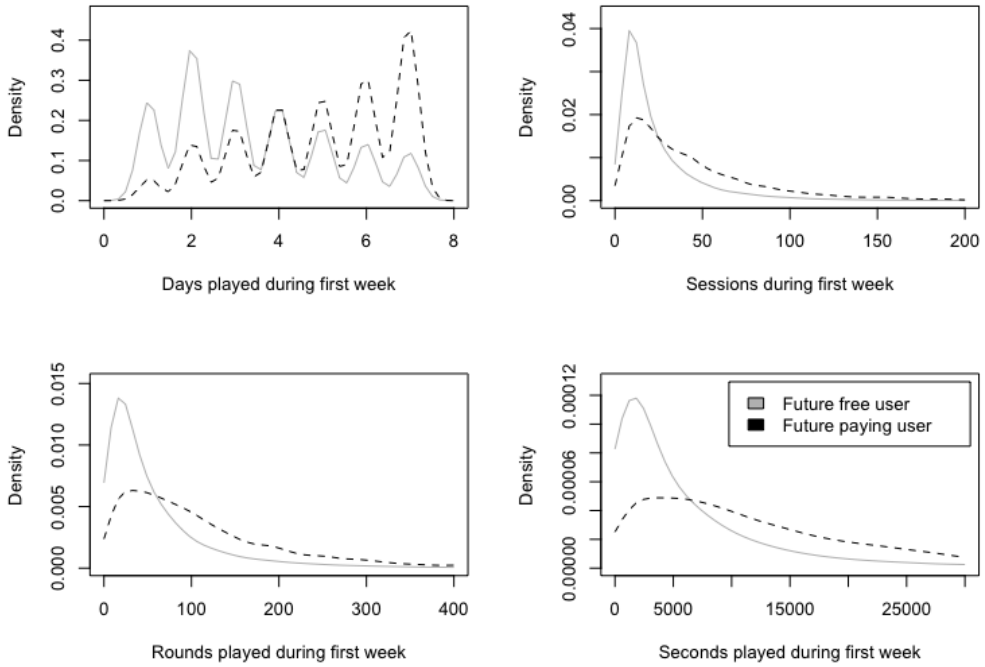
The paper proceeds by presenting relevant conceptual background before providing details on the used methods and results. It closes with a concluding discussion of findings.

4.2 Conceptual Background

4.2.1 Demand prediction in freemium settings

The prediction of future demand has received wide attention in marketing research. In a number of settings, future demand can be captured in the notion of customer lifetime value (Berger and Nasr 1998; Reinartz and Kumar 2003; Gupta et al. 2004) and seminal studies have explored stochastic models of consumer behavior (Schmittlein et al. 1987; Fader et al. 2005) or regression approaches (Malthouse and Blattberg 2005; Donkers et al. 2007; Ekinci et al. 2014) for its prediction. Historically, practitioners seem to have favored simple cross-tabulation and RFM-based (recency/frequency/monetary value) techniques (Verhoef et al. 2003). More recently, reports by data scientists in electronic commerce propose RF algorithms as the method of choice (Vanderveld et al. 2016; Chamberlain et al. 2017). Mobile app datasets are similar to clickstream data commonly available in electronic commerce (Bucklin and Sismeiro 2009) and RFs have indeed been successfully applied to the prediction of purchase decisions in freemium mobile apps (Sifa et al. 2015). Their strong per-

Figure 4.2: Behavioral similarity of future free and paying users

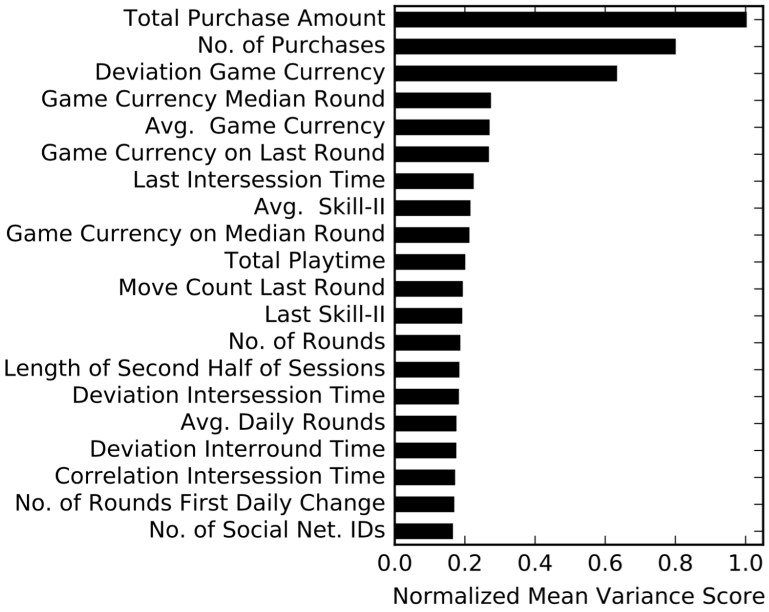


Notes: Distributions of user behaviors during the first week of app use after download, for future free and premium users. Behavioral outcome level on the x-, probability density on the y-axis. While future premium users are more engaged with the app on average, the distributions overlap substantially.

formance in consumer behavior prediction more widely has also been shown, e.g., Lemmens and Croux (2006) or Coussement et al. (2017).

The mentioned applications of RFs (Lemmens and Croux 2006; Vanderveld et al. 2016; Chamberlain et al. 2017) follow an approach akin to a machine learning angle that does not derive methods from explicit consideration of consumer behavior. While less routed in behavioral axioms and oftentimes less effective at providing conceptual insight, such an approach has merits in being agnostic to input data format (Zhang et al. 1998) and assumptions on the data generating process. New data sources that become available over time can be flexibly added to the models (see Section 4.3.3) and predictions can be obtained from diverse sets of input data that may include no or limited accounts of purchase behavior. This flexibility is beneficial to demand prediction in mobile apps where categorical device information, behavioral data from product use, and purchase information all help predict con-

Figure 4.3: Importance of features/variables for prediction of future premium demand



Notes: Variable importance visualization as derived from Random Forest predictor. See Sections 4.3.2, 4.3.3 and 4.3.4 for details.

sumer purchase behavior (Sifa et al. 2015; Runge et al. 2016; also see Figure 4.2). It is further particularly useful in freemium settings more broadly where many future purchasers have not made a purchase at time of prediction and the share of overall premium purchasers is relatively small, leading to class imbalance.

The conversion profiles in Figure 4.1 highlight this: A week after app download (which is the maximum amount of data we make available to predictors in this study, see Section 4.3.3 for a more detailed outline of the underlying reasoning), only about a third of the final share of premium purchasers – that the profile converges to – has materialized. And the final share of premium purchasers is relatively rare with only a low single digit percentage of initial adopters ever making a purchase. Figure 4.3 further shows variable importance as derived from the RF learner: Non-purchase related variables surface close to the top, corroborating their value to the prediction of future premium demand in freemium apps.

In summary, the prediction of future demand in freemium settings is complicated by the following characteristics:

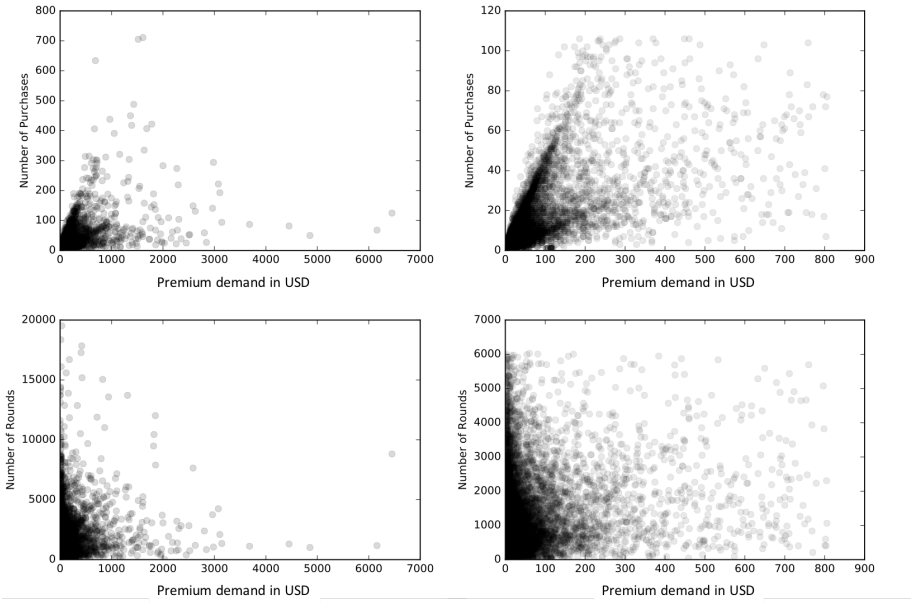
1. Varied behavioral data, both purchase and non-purchase related, are relevant to the prediction (Sifa et al. 2015, Figures 4.3 and 4.4 on the next two pages).
2. As commonly only a small share of users ever purchases a premium upgrade, the prediction target is strongly imbalanced (Weiss 2004).
3. Behavioral outcomes are characterized by discontinuities, limiting the effectiveness of linear learners (Dawes and Corrigan 1974).
4. For the concrete managerial problem studied here, as managers want predictions as early as possible after users' download of an app, purchase histories are short or non-existent.

There however also are benefits to be considered towards a solution for the identified managerial problem: Freemium models can commonly be found in digital settings where firms can log users' digital footprint in databases in a fully automated fashion, providing a large amount of data for estimation. Together with recent developments in computational power, this availability of labeled data makes the application of a data-driven approach attractive. We turn to such an approach. The remainder of this Section motivates our choice in more detail.

4.2.2 Neural networks and choice prediction

Artificial NNs are computing systems inspired by the structure of biological brains (McCulloch and Pitts 1943; West et al. 1997). They are learning frameworks that can accommodate diverse other learning techniques and connect a set of input nodes, the input layer, to a set of output nodes, the output layer (Kumar et al. 1995; Briesch and Rajagopal 2010; more on this in Section 4.3.4, also see Figure 4.6 on page 89). Commonly, a simple NN will have one layer of "hidden" nodes between the input and output layer (Hu and Tsoukalas 2003). Together with other key properties of the NN, i.e., its connectivity (fully connected, feedforward, recurrent) and its nodes' activation function (hyperbolic tangent, logistic, sigmoid), these layers (number of layers, number of nodes per layer) establish the network's topology.

Figure 4.4: Scatter plots of premium demand versus purchases and played game rounds



Notes: Scatter plots of premium demand in USD (essentially overall spending of a user on in-app purchases, for details see Section 4.3.1) versus purchases (top panels) and played game rounds (bottom panels) over the whole observation period; full data on the left, zoomed in on users with premium demand less than 900 USD (99th percentile) on the right. Each dot reflects a user. Many different behavioral profiles lead to similar premium demand, and many similar behavioral profiles lead to different premium demand, indicating diversity in choice paths and non-linear associations in user behavior. A large part of users clustering around the y-axis in the bottom charts play thousands of game rounds, but never make a purchase. Other users spend hundreds of dollars and only play a few hundred rounds.

Deep-NNs are networks with several hidden layers (LeCun et al. 2015). The high flexibility of NNs', particularly deep ones', structure is what awards them with the universal approximation property, i.e., their ability to approximate most classes of linear and non-linear continuous functions to arbitrary precision (Hornik 1991). This property in turn enables them to predict consumer behavior outcomes that show discontinuities for similar levels of certain predictor variables (West et al. 1997, Figures 4.2 and 4.4).

Interestingly, while NNs have found wide application to forecasting more broadly (Zhang et al. 1998; Lessmann et al. 2015), applications of NNs to the prediction of future demand in marketing are somewhat sparse to date (Müller-Navarra et al. 2015), and applications of deep learning are virtually non-existent. Hu and Tsoukalas (2003) consider a simple NN topology with one hidden layer to investigate con-

sumers' choice of communication modes in telephony. Kim et al. (2005) apply NNs in combination with genetic algorithms to the prediction of auto insurance demand and find the NN approach to be beneficial when managers have clear decision criteria. Glady et al. 2009 find cost-sensitive classification techniques to be more profit-optimal for the prediction of churn via customer lifetime value and use a NN as part of their benchmarking methods. Their NN's architecture however does not receive explicit attention and is not "deep," i.e., it does not have several hidden layers. Briesch and Rajagopal (2010) consider applications of NNs in consumer research. The network in their regression tasks has one hidden layer with three nodes (note that the final network of the present study has four hidden layers with hundreds of nodes, see Figure 4.6 for details). The network in their non-linear principal components analysis application is a rare exploration of a deep architecture, however applied to the identification of behavioral constructs rather than future demand. Moro et al. (2015) explicitly consider NNs to predict the demand for long-term bank deposits. Their NN has one hidden layer and is hence not deep either. One reason for the scarcity of deep learning-related studies is deep-NNs' high need for computational resources and large samples. This limitation has however been alleviated in recent years due to advances in computation (LeCun et al. 2015) and digitization enabling the collection and storage of large and dense labeled datasets (Lambrecht et al. 2014).

Kumar et al. 1995 compare NNs and logistic regression for the prediction of managerial choices and highlight both advantages and challenges. West et al. (1997) study NNs in the context of consumer choice prediction and find them to be particularly useful for their ability to capture non-linearities in consumer behavior. Figure 4.2 shows that many consumers in freemium settings use a product to a similar extent as a future paying user, but never make a purchase. This suggests that they use decision rules where no price can entice them to purchase a premium upgrade even if the expected complementarity from said upgrade is large as they use the product a lot.

To leverage the availability of dense records of consumers' digital footprints, and to overcome complications 1, 3 and 4 mentioned in the previous Section 4.2.1, the present study applies a NN approach to the prediction of future demand in a freemium mobile app. Over and above extant literature, the NN topology considers several hidden layers (and is hence "deep") to maximize flexibility in learning discontinuous outcomes from consumer choices. The study further exposes both the learning process – an extensive hyper parameter grid search as described in Section 4.3.4 – and the final network's topology in detail. The best performing NN has four hidden layers, for a total of six layers with input and output layer. The authors additionally introduce a synthetic oversampling technique, speaking to complication 2 in Section 4.2.1, and further aiding the NN's performance. The next section provides background on this technique.

4.2.3 Synthetic oversampling

Relevant behavioral prediction targets in clickstream datasets are often characterized by class imbalance (Lessmann 2004; Weiss 2004), e.g., the share of converting users is small compared to all users exploring an electronic commerce website (Moe 2003; Moe and Fader 2004). This imbalance is mostly caused by rare classes rather than rare events (Weiss 2004) as the number of observed entities is usually large. Commonly, correct prediction of rare entities entails the higher value proposition. This is certainly the case for premium users in freemium settings, but also for conversions on a website or an ad, and the prediction of user disengagement (Coussement et al. 2017). Conventional supervised machine learning methods tend to lean towards the majority class especially when dealing with highly unbalanced datasets (Weiss 2004). The dataset used here is highly unbalanced as is representative of freemium datasets more broadly: Premium users commonly only make up a low single digit percent of the overall user base (Lee et al. 2017; Runge et al. 2014, 2016). In the dataset used in this study, the share of premium users amounts to 2.1% at

the end of the 360-day observation period. It is generated by the app that is behind the median conversion profile in Figure 4.1 on page 73.

To aid NNs in not overfitting, we choose a synthetic oversampling technique (SMOTE) proposed by Chawla et al. (2002). SMOTE augments datasets in a self-reliant manner and can be adapted to behavioral datasets (Sifa et al. 2015, 2018). During the training phase, it creates synthetic instances that are random convex mixings of actual instances in the minority class (Chawla et al. 2002) to regularize the prediction models to avoid overfitting and to enable learning of structures representing minority entities. In some ways, SMOTE resembles distortion-based model regularization techniques (Bishop 1995; Vincent et al. 2008).

In the studied freemium setting, it generates additional synthetic premium users from actual premium users' data records by randomly mixing a user's count and numeric features with these of one of its k nearest neighbors and inheriting the original user's categorical features.[3] By oversampling the class of premium users who spend money and have non-zero future premium demand, it reinforces the signal contained in these app users' "touchstream" data. This enables models to pick up on choice paths and behavioral profiles relevant to the generation of the prediction target. At the same time, the convex mixing of actual users' behavioral profiles does not just duplicate existing information, but adds noise, too. This added noise supports NNs in avoiding overfitting to spurious associations in the data (Bishop 1995; Vincent et al. 2008).

4.2.4 Conceptual expectations

Before formally presenting the prediction problem, the data, and our estimation approach, we wish to summarize our expectations towards the empirical analysis based on the background provided in the previous sections: First, we expect the

[3] We refer the interested reader to Sifa et al. (2018) who present a detailed exposition of SMOTE and how it can be adapted to behavioral data. We use their implementation in our analysis.

search across different network topologies to yield a deep architecture for the best performing NN. Second, as both NNs and RFs are able to learn non-linearities, we expect them to outperform linear regression (LR) when it comes to making actual zero predictions for users who use the app, but end up never making a purchase. Third, we expect learners to be better able to identify future premium customers when oversampling is applied. Fourth, as the dataset is large and we do not consider computational constraints, we expect NNs to ultimately be best able to detect future premium customers, particularly when combined with synthetic oversampling.

4.3 Method

4.3.1 The prediction problem

We define overall premium demand as the cumulative gross amount spent by a user until day 360 (about a year) after adoption of the app. This approach is in line with current managerial practice in the app economy (Seufert 2013). We focus on users' spending on in-app purchases (IAPs) as advertising revenue is of low relevance in the studied setting (Ghose and Han 2014) and the studied freemium app did not include advertising at the time of data collection. We further disregard cost as it is virtually zero at the margin (Lambrecht et al. 2014) and, in line with marketing practice in this environment, we do not apply discounting (Seufert 2013). To have a consistent prediction target outside the input data range (see Section 4.3.3), we aim to predict users' premium demand between day eight and 360 after product adoption. Denoting cumulative spend on IAPs of a user along days after app download as $y_{day\,j}$, we can formalize the prediction problem as:

$$IAP_{future,i} = y_{day\,360,i} - y_{day\,7,i} = f(X_{t,i}) \tag{4.1}$$

$IAP_{future,i}$ hence represents the total amount spent on IAPs by user i between day eight and 360 after their respective app download. A user's overall premium demand is given by the sum of $IAP_{future,i}$ and $y_{day\,7,i}$. $X_{t,i}$ captures all available click-stream data for user i at the time of prediction t, with $t \in \{$*app download*, *one day*, *three days*, *one week*$\}$ as described in Section 4.3.3 and shown in Table 4.1. Finally, f is the function that we want to approximate with our learners to yield user-level predictions of future premium demand from the various input datasets – similar in spirit to Malthouse and Blattberg (2005). We will approximate it using a NN, RF and LR learner by minimizing the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{day8-day360,i} - y_{day8-day360,i})^2} \tag{4.2}$$

We choose RMSE over other error measures as it is commonly applied and has the property to weigh larger deviations more strongly, i.e., it penalizes wrong predictions for higher premium demand relatively more. We further choose to measure premium demand as a continuous USD amount rather than as a binary indicator to keep granularity high and enable learners, particularly the NN, to observe diverse choice paths leading to differing outcomes. In a binary classification framing, the breadth of possible outcomes would be drastically reduced, rendering the application of learners that can discern various decision rules and paths less attractive.

4.3.2 Sample

The sample used in analysis comprises detailed tracking logs of 197,665 adopters of a freemium mobile gaming app. Because of the sensitive revenue data made available for this study, the data sponsor chose to remain anonymous. The app builds on game levels that consist of visual puzzles and are arranged on a map where players have to solve a level to access higher levels – representative of freemium gaming apps such as Candy Crush Saga (Levitt et al. 2016). Users can make premium purchases

Table 4.1: Overview of input data

Data layer	*Notation*	*Contained input variables / features*
Background	X_t with $t \in \{app\,download\}$	Country segment, device type, operating system, acquisition type (paid marketing versus organic)
Behavioral (generic)	X_t with $t \in \{one\,day, threedays, oneweek\}$	Number of sessions, number of rounds, number of active days, number of purchases, total purchase amount, total playtime, number of social network connections
Behavioral (product-specific)		Total score, number of lives, amount of game currency, number of cleared puzzle elements, difficulty level, game type, moves count, level outcome, skill measure I, skill measure II, skill measure III
Temporal		Total inter-session time, total inter-round time, time between daily first and last session, inter-day time distribution, inter-session time distribution
Composite		For per-day and per-session behavioral features: Correlation coefficients over time; first order trends; maximum, mean, median and deviation over time Per game type and difficulty level: Activity ratios and entropy

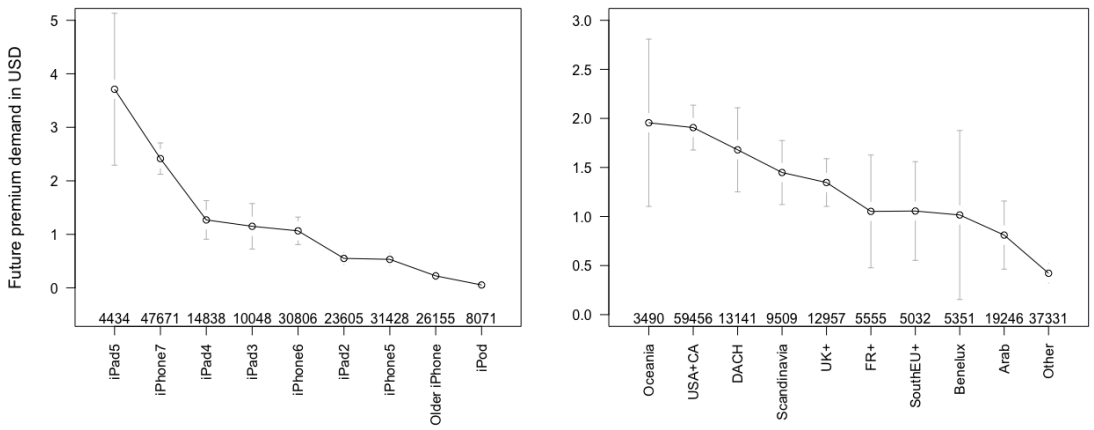
Notes: Background data are available right at app download, behavioral data accumulate as users use the app after download. Temporal and Composite features/variables are engineered to enhance the data's predictive power.

via IAPs that afford them with in-game currency that can then be used to enhance gameplay and unlock paywalls (Runge et al. 2016). It had been played by close to 100 million users across Apple's, Google's and Facebook's app marketplaces as of January 2018. The data at use here originate from users on Apple's distribution platform.

4.3.3 Datasets

Managers' goal is to have estimates of users' future premium demand early on, to take informed action as soon as possible after users' adoption of the freemium product or service (Malthouse and Blattberg 2005; Seufert 2013). The earliest point in time when user data for a prediction are available to app publishers is at app download – this is when categorical background data in the form of geolocation and device information are collected (Bucklin and Sismeiro 2009). Such predictions can inform

Figure 4.5: Future premium demand by device and country segments



Notes: Categorical meta data observed at app download (country and device characteristics) is relevant to the prediction of premium demand: Average future premium demand in USD with 90% confidence interval for device and country segments. Segment size noted above the x-axis.

marketers' interaction with users right from their first moment of product/service use.[4]

Behavioral data – e.g., logins to the app, played game rounds, purchases made (see Figure 4.2 on page 76) – then are recorded in the company's tracking logs as consumers use the product, increasingly make choices and reveal their behavioral patterns (Bucklin et al. 2002; Bucklin and Sismeiro 2009). Some users will disengage early on, making future revenue contributions on their end highly unlikely while other users will effectuate first premium purchases (Moe 2003; Moe and Fader 2004).

Similar in spirit to prior studies that assess the sensitivity of customer classifications to different amounts of data (Heilman et al. 2003), we predict future premium demand from datasets that comprise different amounts of behavioral data and are available at different points in time after product adoption. As behavioral data in apps display a weekly periodicity (see Figure 4.1 on page 73), we opt for increments in behavioral data up to a week (Heilman et al. 2003). This yields the following datasets for prediction:

[4]To exemplify the predictive value of the geolocation and device information available at app download, Figure 4.5 shows future premium demand for different device and country groups as used by the marketing managers and analysts of the data sponsor.

1. Background data only ($X_{app\ download}$; see Table 4.1);
2. Background data and behavioral data from users' first day ($X_{one\ day}$);
3. Background data and behavioral data from users' first three days ($X_{three\ days}$);
4. Background data and behavioral data from users' first week in the product ($X_{one\ week}$).

We include the complete digital footprint logged by the company's databases in the datasets used for prediction. Table 4.1 presents an overview of the variables contained in these datasets. Before training prediction models, we enriched data by adding two additional layers; a temporal layer that captures inter-event time distributions and a composite layer that contains correlations, trends and deviations of features over time.

We further apply synthetic oversampling. We rely on the adaptation of SMOTE to behavioral datasets presented in Sifa et al. (2018). SMOTE is applied to datasets 2, 3 and 4 above that contain behavioral data, yielding a total of seven datasets. We synthetically oversample the minority class of 2.1% premium users by factors of 5, 10 and 20 to generate shares of premium users close to 10, 20 and 30% which are considered thresholds for (strong) imbalance in datasets (Weiss 2004). We find 5 to provide the best results (note that oversampling is only applied on training and not on testing data). The oversampled datasets that we use to train models hence have a share of 9.5% premium users instead of the 2.1% in the baseline dataset.

4.3.4 Estimation and learner implementation

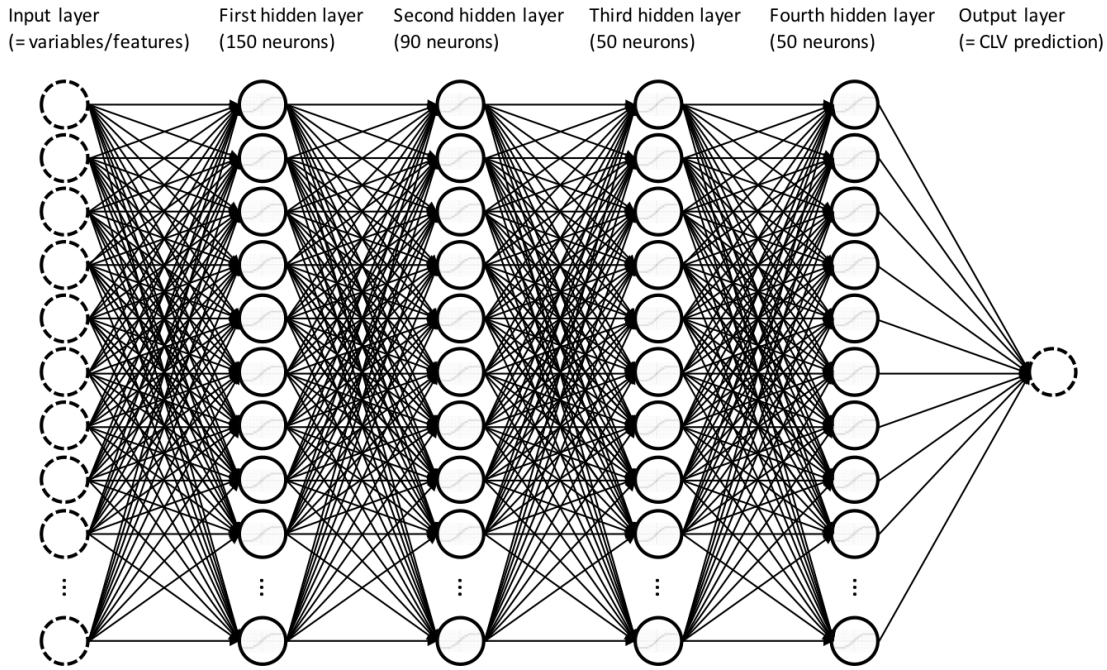
Marketing models are often estimated and evaluated based on a time-separated holdout group (Fader et al. 2005; Dziurzynski et al. 2012). This approach is routed in the firm's problem to generate predictions about future behavior(s) of a set of current customers at current time. In the estimation and evaluation of machine learning models, researchers often apply cross-validation for estimation of the model on the training sample (Angermueller et al. 2016; Yadav and Shukla 2016). K-fold

cross-validation splits the sample into k random folds and uses k-1 for training and one fold for testing until each fold has been used for testing. Another common approach splits data into a training, validation and holdout set (Surrette 2019). A random training sample is used to estimate the model, the – mostly small – validation sample is used to validate the estimate and re-estimate the model if e.g., the difference in estimate between training and validation sample is large, and the holdout sample is finally used to cleanly evaluate the overall model. We adopt this approach for our grid search of the best hyper parameters.

We implement NNs as fully connected feed-forward multilayer perceptrons (Hornik 1991). The grid search spans between one to four hidden layers with 50 to 150 neurons (in steps of 10) each and three different activation functions for neurons – hyperbolic tangent (tanh), sigmoidal (LeCun et al. 1998) and rectified linear unit (ReLU – Glorot et al. 2011), yielding a total of 4 x 11 x 3 = 132 hyper parameter combinations to consider. We apply backpropagation based on a batched gradient descent optimizer with an adaptive learning rate (Kingma and Ba 2014) to learn the weights of neurons. We further apply dropout regularization (Srivastava et al. 2014) in addition to SMOTE to avoid overfitting and improve generalization (Bishop 1995; Vincent et al. 2008). Dropout regularization randomly blinds different portions of the weights during training (Srivastava et al. 2014). Estimations for NNs were implemented on TensorFlow using the Python library Keras. RF and LR were implemented as available in the package Scikitlearn in the programming language Python. SMOTE was adapted from Sifa et al. (2018). For RFs, we tested configurations with 50 to 250 trees. The only parameter tuned for LR is the intercept, otherwise the resulting model coincides with the model identified on the full training sample. Each individual learner is optimized by minimizing root mean squared error (RMSE) of predictions (see Equation 4.2 on page 84). We choose RMSE over other error measures to penalize large deviations.

It should be noted that our approach focuses on cross-sectional estimation and

Figure 4.6: Visualization of the final network's topology



Notes: Topology of the best performing NN as identified in a grid search on the most complete dataset containing background data and a week of behavioral observations. Note the feedforward structure where neurons in one layer have no lateral connections. Layers are fully connected and neurons have a hyperbolic tangent activation function.

validation as mentioned above. To be effective for temporally separated use, i.e., estimation based on current users and prediction on future users, this approach requires the assumption of stationarity of the data generating process over time. This assumption is often reasonable and prediction models' robustness can be increased by either devising an explicit temporal validation of estimations or by continuously retraining the model with newly arriving data and, e.g., overweighting more recent observations in model estimation.

4.4 Results

We split the data 50/10/40, with 50% of data being used for training, 10% for validation of model performance and 40% as a holdout sample. The 10% validation sample is used to validate grid search results obtained on the 50% training sample

and to tune hyper parameters until results on training and validation sample closely coincide. Results presented in this section are obtained from ten draws without replacement on the 40% holdout sample. These bootstraps are used to generate confidence intervals around RMSE and hit rate results and provide an indication of statistically significant differences between results.

We use and report RMSE mainly for diagnostic reasons. To evaluate the learners as to their managerial usefulness, we focus on the different algorithms' hit rates[5] as derived from a sorting of users by their demand prediction (Malthouse and Blattberg 2005; Lemmens and Gupta 2020). We sort users by their predicted future premium demand and consider different ratios of the upper end of this ranking. The hit rate then informs us how many actual premium users we correctly predict when we consider these different ratios of the prediction-sorted sample (Malthouse and Blattberg 2005).

The grid search – for details on the grid that we searched across see Section 4.3.4 – identified a NN topology as shown in Figure 4.6 as best performing in terms of achieving lowest RMSE. The best performing NN has four hidden layers and a hyperbolic tangent activation function (see Figure 4.6). This result speaks to our first conceptual expectation formulated in Section 4.2.4: A deep-NN architecture appears to outperform NNs with a lower number of hidden layers. In the following, we first present what variables are relevant to the prediction based on the RF learner and RMSE results before centering in on hit rates, i.e., how well different learners do on different datasets in the identification of future paying customers.

4.4.1 Variable importance and RMSE

Figure 4.3 – in line with existing literature (Fader et al. 2005; Malthouse and Blattberg 2005; Voigt and Hinz 2016) – shows that purchase-related variables surface

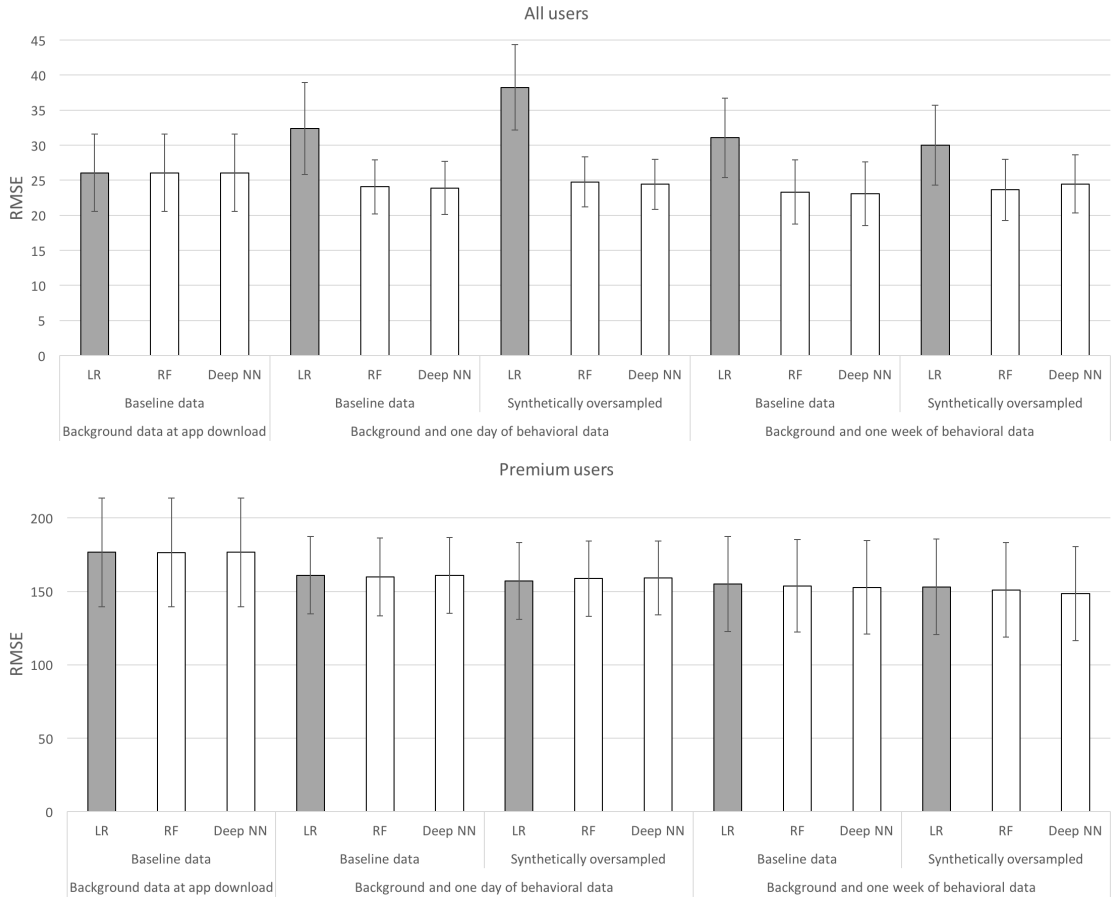
[5]We define hit rates equivalently to recall as the number of true positives (here: correctly identified future paying customers) over the sum of true positives and false negatives (here: the total number of future paying customers).

as most important for prediction of premium demand, but usage-related features such as number of rounds, inter-session time and measures of users' skill also surface as relevant predictors (Moe and Fader 2004; Sifa et al. 2015). Figure 4.7 depicts average RMSE with 95% confidence intervals for all learners and five different input datasets. Confidence intervals are obtained from ten draws without replacement on the holdout sample. We report RMSE for two distinct user groups: All users and premium users only. While these groups are not known at prediction time, the ex-post segmentation allows us to understand how the relative performance of the methods differs between non-paying and paying users.

The key difference that emerges between learners speaks to the second conceptual expectation that we formulated in Section 4.2.4. LR underperforms for predictions on the full user base (upper panel of Figure 4.7). The RMSE for LR spikes when behavioral data are added compared to predicting only from categorical background data, both with and without oversampling. This underperformance is driven by LR's inability to successfully learn the discontinuity that users display at the zero-price point. It becomes salient when behavioral data are added as usage-related data make premium and non-premium users look similar in a linear perspective as illustrated in Figure 4.2. Figure 4.4 also depicts an angle to rationalize this: Many non-premium users play just as many or more rounds than premium users. LR cannot predict the discontinuity in consumer behavior resulting from a decision rule of the sort "I will not spend money regardless how many rounds of this game I play." In the words of West et al. (1997), "neural network models can offer significant improvement over traditional statistical methods because of their ability to capture nonlinear relationships" (p. 1).

The lower panel of 4.7 further substantiates this insight. When we remove the zero-point non-linearity and only consider RMSE in the segment of premium users, the underperformance of LR disappears and all of LR, NN and RF perform similarly well. LR's underperformance is hence routed in non-zero predictions for actual zero

Figure 4.7: RMSE results for different input datasets across for all users and for premium users only



Notes: RMSE results (mean with 95% confidence interval) for different learners as obtained on the holdout sample for three different datasets with and without synthetic oversampling. Note that synthetic oversampling is only applied to behavioral datasets as described in Section 4.3.3. The upper panel shows RMSE for all users, the lower panel for premium users. Results for LR are highlighted in grey to emphasize how it underperforms for predictions across all users, but not when only considering premium users. The reason for this is LR's inability to learn the discontinuity at the zero-price point: Some users will never spend money regardless how much they use the app, LR cannot learn this non-linearity in consumer behavior while RF and deep-NN can (Einhorn 1970; West et al. 1997).

premium demand users (that are no longer included in the results in the lower panel). Figure 4.7 further suggests that RF can learn such nonlinearities in our context similarly well as a deep-NN.

Overall, there are no differences in RMSE between RF and NN (upper panel), but both perform better than LR. Mild improvements in RMSE become apparent for an increasing amount of behavioral data, the application of oversampling, and for the Deep-NN over the other learners when we consider premium users only and

Table 4.2: Complete overview of hit rate results

Input data	SMOTE	Method	Top 5%	Top 10%	Top 15%	Top 20%	Top 25%	Top 30%
App download	No	LR	14.28%	26.51%	36.00%	46.11%	52.00%	56.77%
		RF	14.19%	23.50%	34.95%	44.55%	50.03%	58.37%
		Deep-NN	*14.15%*	*25.43%*	*35.89%*	*45.56%*	*52.41%*	*56.89%*
One day	No	LR	25.76%	41.50%	54.90%	64.14%	70.81%	75.50%
		RF	18.45%	31.14%	40.64%	50.23%	57.31%	63.63%
		Deep-NN	29.19%	44.01%	56.87%	65.31%	71.82%	76.73%
	Yes	LR	28.23%	44.58%	56.97%	65.56%	72.73%	77.09%
		RF	21.85%	34.28%	45.37%	56.33%	65.19%	71.14%
		Deep-NN	*29.84%*	*45.30%*	*56.99%*	*66.23%*	*72.47%*	*77.84%*
Three days	No	LR	36.55%	50.63%	60.63%	68.31%	73.69%	77.54%
		RF	28.62%	39.74%	49.48%	57.72%	63.98%	70.38%
		Deep-NN	40.31%	54.05%	63.16%	70.62%	76.70%	80.84%
	Yes	LR	36.48%	49.87%	59.61%	67.42%	74.20%	78.23%
		RF	31.35%	42.73%	53.23%	62.64%	69.58%	75.04%
		Deep-NN	*40.78%*	*53.55%*	*63.39%*	*70.88%*	*77.16%*	*81.32%*
One week	No	LR	40.94%	52.23%	61.10%	68.26%	72.86%	76.33%
		RF	34.62%	45.14%	53.85%	60.42%	67.37%	75.45%
		Deep-NN	43.22%	55.93%	64.80%	71.09%	75.91%	80.07%
	Yes	LR	38.79%	52.09%	62.32%	69.06%	73.86%	78.91%
		RF	37.56%	47.95%	56.51%	64.50%	71.14%	76.60%
		Deep-NN	*42.73%*	*56.20%*	*65.59%*	*72.46%*	*78.58%*	*83.24%*

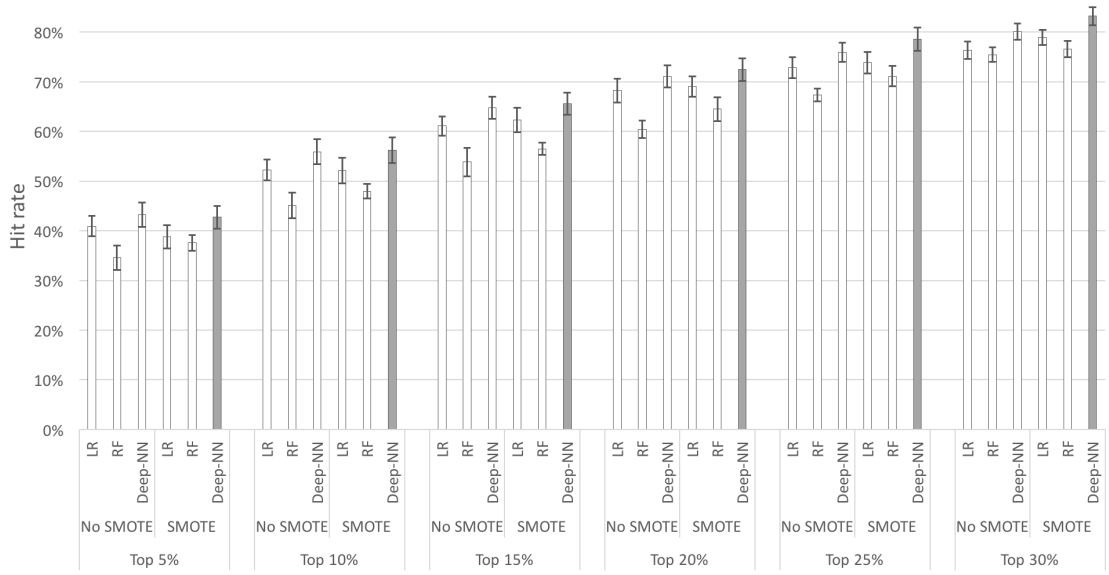
Notes: Mean hit rates for different ratios of the demand prediction-ranked sample and for the different datasets as described in Section 4.3.3, obtained from ten random draws without replacement on the 40% holdout sample. Reading help: The bottom entry in the last column tells us that, when targeting the top 30% of the demand prediction-ranked sample, the deep-NN predictor trained on synthetically oversampled data will reach 83.24% of future premium users. Deep-NN results on oversampled data are highlighted through italics as they provide the overall best hit rate. Figures 4.8 and 4.9 visualize a subset of the results shown here to highlight the results' sensitivity to the size of the top x% considered for the hit rate and to the amount of behavioral data included in the prediction.

move towards the right of the lower panel in Figure 4.7. Average future premium demand is \$1.17 across all users and \$55.01 for premium users. The square percentage error of predictions over actuals expressed as RMSE divided by actual future premium demand in the respective user segment is hence at around 2000% for all users and 250% for premium users. These high error numbers suggest that mobile app clickstream data only poorly capture the actual data generating process of (future) premium demand in freemium settings which is in line with findings by Bucklin and Sismeiro (2009) for clickstream data (more on this in Section 4.5).

4.4.2 Hit rates: Predicting future paying customers

In practice, marketers often rely on a sorting of users in terms of future value to inform interactions with consumers (Malthouse and Blattberg 2005; Lemmens and

Figure 4.8: Hit rates for different ratios of users sorted by demand predictions

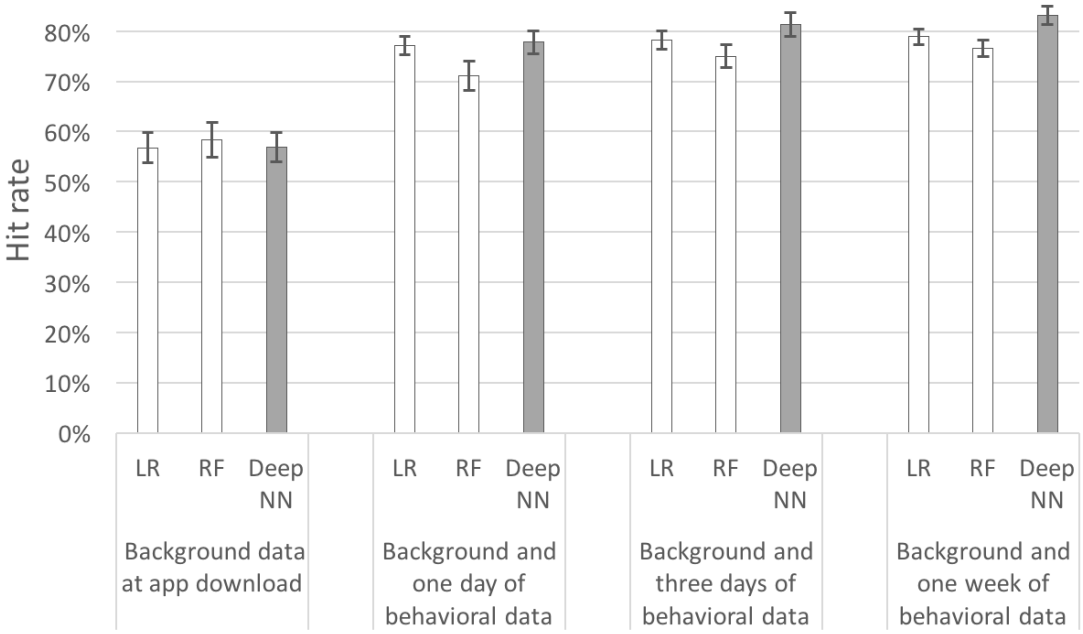


Notes: Mean hit rates with 95% confidence intervals as obtained from ten random draws without replacement on the 40% holdout sample. Synthetic oversampling was not applied to the background data available at app download, hence the figure only shows results for non-oversampled data there. Underlying values are listed in Table 4.2. Results for the deep-NN combined with synthetic oversampling are highlighted in grey. It performs best on average and significantly so for larger top x percentage shares of the prediction-ranked user list.

Gupta 2020) and the allocation of marketing budget to different channels where channels with high shares of future premium customers may receive more resources (Blattberg and Deighton 1996). To assess the effectiveness of the different learners in providing this decision support, we sort users based on the individual-level demand predictions from high to low. We use this sorted user list to analyze the hit rate (or recall) for actual premium users along the obtained ranking.

Figure 4.8 presents these hit rates for different top ratios of the sorted list. As we can see from a comparison of hit rates with and without synthetic oversampling, all learners benefit from SMOTE on average. This insight speaks to our third conceptual expectation formulated in Section 4.2.4: Learners are better able to identify future premium customers when oversampling is applied. Figure 4.8 further shows that, when a larger share of the user prediction-ranking is considered, and particularly when synthetic oversampling is applied, the deep-NN outperforms the other methods (as indicated by the 95% confidence intervals around the mean hit

Figure 4.9: Hit rates for top 30% of users for different datasets with synthetic oversampling



Notes: Mean hit rates with 95% confidence intervals as obtained from ten random draws without replacement on the 40% holdout sample. Underlying values are listed in Table 4.2.

rate for different ratios). This confirms our fourth and last conceptual expectation in Section 4.2.4 that a deep-NN will be best able to identify future premium customers in this setting with diverse choice paths and non-linear decision rules.

With a week of behavioral data and oversampling, the deep-NN is able to identify 83.2% of actual future premium users when we consider the top 30% of the prediction-ranked user list (see Table 4.2 that summarizes all hit rates underlying Figures 4.8 and 4.9). The same figure is at 78.9% for LR and 76.6% for RF. This difference is statistically significant per Figure 4.8. Using a deep-NN and SMOTE, marketers are able to identify future premium users more accurately and can use this information to inform interactions in customer relationship management. We will discuss concrete examples in Section 4.5.1.

Figure 4.9 shows results on oversampled data for different amounts (from none to a week) of behavioral data per the different input datasets described in Section 4.3.3.

It corroborates that more behavioral data improve prediction performance for all learners (Heilman et al. 2003). It further shows that the outperformance of the deep-NN over the other learners materializes as more days of behavioral data are added and hence more information describing consumer choices are made available to learners. This observation seems sensible as the deep-NN is expected to more fully develop its universal approximation property the more data are available to it (West et al. 1997; LeCun et al. 2015). Table 4.2 summarizes all hit rates underlying Figures 4.8 and 4.9.

4.5 Discussion

In freemium settings such as the app economy (Arora et al. 2017; Ghose and Han 2014), a small share of users drives revenue and co-finances free provision of the product (and development of new products) for all users (Bapna et al. 2017). In the app studied here, only 2.1% of users purchase a premium upgrade which is representative of such settings (Shi et al. 2015; Lee et al. 2017; Gu et al. 2018). The existence and retention of premium users is essential to firms' survival (Shi et al. 2019). Marketers wish to identify these users early after their adoption of an app, to tailor and target marketing initiatives for retention of existing and acquisition of new users of this kind.

In speaking to this quantitative managerial problem, the present study leans into deep learning that has achieved breakthrough success in other fields (LeCun et al. 2015; Angermueller et al. 2016). (Deep) NNs suggest themselves for application in freemium settings as they can flexibly learn non-linear decision rules that cannot be discerned by linear learners (West et al. 1997). Such decision rules originate from the documented difficulty to remove certain consumers from a zero-price point (Shampanier et al. 2007). While these consumers use a freemium product in a similar way as a premium user (see Figures 4.2 and 4.4), they keep "sampling" for free and

never spend money (Bawa and Shoemaker 2004). The authors compare a deep-NN's ability to identify future paying customers among app users to the ability of a RF and a linear learner. They further present a synthetic oversampling approach (Chawla et al. 2002) to reduce class imbalance that additionally has a regularization effect (Bishop 1995; Vincent et al. 2008). Its application benefits all learners and particularly the NN that emerges as a capable detector of future paying customers.

This section discusses how the presented methodology can assist marketers before highlighting limitations of the presented work and pathways for future research.

4.5.1 Informing marketers' interactions

If the firm was to blindly target users without data-driven decision support, it could reach 30% of future premium users by targeting a random 30% of users. Our deep learning-based approach can target 83.2% of actual future premium users by targeting the top 30% of prediction-sorted users (see Table 4.2 on page 93). This clearly is an improvement (of 53.2 %-points) over the naive baseline, but also outperforms both the linear and RF learners by more than four percentage points (see Table 4.2 on page 93 for details). Speaking to substantive applications, these results can be leveraged to inform marketers' interactions in several areas (Blattberg and Deighton 1996; Berger and Nasr 1998; Heilman et al. 2003):[6]

- Customer acquisition: Marketing managers like to evaluate newly acquired users through digital display or search engine advertising (Zenetti et al. 2014; Guhl et al. 2016) as to their expected revenue contribution (Seufert 2013). This practice is used to ensure that advertising budget expensed on acquisition of new users does not exceed these users' expected future revenue contributions (Blattberg and Deighton 1996; Seufert 2013).
- Customer service: Future paying customers can be prioritized in customer

[6]These examples are intended to illustrate possible applications. We however cannot and do not intend to make statements as to the causal impact of the discussed applications. Identifying these requires experimentation and/or further analysis.

service. While the wider product may be operated using chat bots (Sivaramakrishnan et al. 2007), high-value prospects can be contacted by a dedicated human customer service agent. Concretely, marketers could decide to handle requests from the top 20% of prediction-sorted users using human rather than machine agents. This ensures that 72.5% of future paying customers receive high-quality responses, but it will reduce workload for human agents by 80% compared to handling all users' requests (assuming requests come in equally across users).

- In-app advertising: Many apps and digital news outlets monetize their user base by exposing them to advertising (Lambrecht and Misra 2016; Halbheer et al. 2014; Ghose and Han 2014). Marketers may choose not to expose future paying customers to advertising to ensure a seamless product experience for them (Calder et al. 2009). E.g., the firm can decide to not expose the top 30% of the prediction-sorted user list to advertising – which ensures that 83.2% of paying customers have a seamless product experience. At the same time, this approach will remove 28.3% of non-paying users from advertising exposure. So, if the firm deems advertising exposure of non-premium users important, it may want to set a different cut-off to define premium customers and, e.g., only consider the top 10% of the prediction-sorted user list as premium customers.
- Promotions and personalized pricing: Marketers are likely to benefit from extending targeted promotional offers that raise awareness for relevant premium upgrades (Zhang and Wedel 2009). Our predictions can inform this targeting, e.g., marketers can avoid giving small and/or low price point offers to high-value prospects. Predictions can further be an input for price personalization policies (Misra et al. 2019).

4.5.2 Limitations and future research

The data at use in this study (see Sections 4.3.2 and 4.3.3) are representative of the datasets generally available to vendors of mobile apps. RMSE results presented in Section 4.4.1 indicate that, despite a large sample and dense behavioral observations, the user-level data generating process is only poorly captured. Bucklin and Sismeiro (2009) state that "to effectively use the information, clickstream data usually needs to be augmented and matched with other sources within the firm" (p. 46). Clickstream data available in the app economy seem to present similar challenges. We propose a technique (SMOTE – Chawla et al. 2002) that augments imbalanced datasets in a self-reliant manner and improves the selected methods' prediction performance. In particular, it increases deep-NNs' ability to learn a ranking of users by future premium demand from the dataset. When facing unbalanced data more broadly, it appears worthwhile to explore the use of (synthetic) oversampling (Chawla et al. 2002). Oversampling can further only be applied to user segments that are deemed particularly valuable. E.g., the firm may choose to only oversample users of very high-value, say the top 20% of paying customers (that account for more than 80% of revenue in the studied setting). This approach can enable learners to detect users in this segment with higher accuracy and may prove to be a valuable avenue for future studies and applications. Another viable avenue for future inquiry may present itself in augmenting (mobile app) clickstream data with psychological measures, e.g., personality-related information (Kosinski et al. 2013; Matz et al. 2017).

Over the years, marketing research has produced workhorse models for the prediction of future demand that are routed in models of consumer choice processes, e.g., "buy-'til-you-die" models (Schmittlein et al. 1987; Fader et al. 2005). It seems promising to explore adaptations of such models to detailed digital footprints of consumer behavior to improve predictive accuracy (Dew and Ansari 2018). Along these lines, stochastic models could be extended to make predictions of future purchases from "free" app use data to compare their performance to the learners presented in

this study. An advantage of such an approach can be its ability to provide conceptual insight on consumer choice processes. Speaking to this, we wish to point to methods to derive conceptual insight from machine learning methods. We use the RF learner to create an overview of variable importance (see Figure 4.3). Such a list of important predictors is useful as it can provide insight into consumer choice processes: E.g., since user skill surfaces as a meaningful predictor of future premium demand (see Figure 4.3), managers can devise in-app tutorials to help users develop their skills, in turn possibly positively impacting their future premium demand. Important predictors can further be used to develop more granular models of consumer choice processes that reach beyond purchase behavior. In this sense, there are avenues for machine learning methods and existing marketing models to beneficially influence each other.

We further wish to note that the implementation of NNs presents challenges. Choosing the right topology – e.g., number of hidden layers, activation function, connectivity between layers – is not an easy feat and requires highly trained experts to successfully leverage NNs' potential. Special attention needs to be placed on avoiding overfitting – as (deep) NNs can so flexibly accommodate diverse functions mapping inputs to outputs, they can easily overfit to associations in the data that do not generalize. Large samples and techniques to avoid such overfitting are hence essential (see Section 4.3.4; Srivastava et al. 2014; Bishop 1995). The training and maintenance of NNs is further computationally expensive. Vast improvements in computational ability over the last decades however alleviate this challenge (LeCun et al. 2015) and researchers can assist analysts in identifying NN topologies and training modes that work well for particular problems. The present study attempts to achieve this: The setup detailed in Section 4.3.4 can be a starting point for analysts and researchers faced with diverse consumer behavior predictions in digital settings. Further, challenges commonly associated with NNs, namely incomprehensiblity and incorporation of prior knowledge, are likely to be increasingly remedied

as researchers are starting to put machine learning methods on more robust statistical foundations and, e.g., enable their use for causal inference rather than mere prediction (Wager and Athey 2018; Alaa et al. 2017).

Finally, it should be noted that data pre-processing can impact predictor performance (Crone et al. 2006; Coussement et al. 2017). Extensive pre-processing can capture non-linearities that a NN will be able to model independently through its hidden layers, but a LR will not be. While the authors handcraft a wide range of input features (see Section 4.3 and Table 4.1) that support algorithms in, e.g., capturing non-linearities, it can be worthwhile to explicitly explore the sensitivity of results to different data pre-processing approaches (Crone et al. 2006; Coussement et al. 2017). A further promising line of inquiry is the use of NNs for the automated encoding of input data (Chamberlain et al. 2017) which can increase the predictive performance of varied prediction models.

Chapter 5

Monetizing Freemium Play: A Practical Evaluation of Pricing Tactics in a Mobile Game

Julian Runge

Abstract

Apps and app stores have become the dominant distribution channel for mobile content. As marginal cost of production and distribution are virtually zero, most apps are priced "freemium." Due to this particular pricing structure and the ubiquitous availability of mobile devices, gaming has seen an unparalleled increase in demand, with mobile gaming now accounting for three quarters of $101 billion in-app purchase revenue and 60% of overall gaming revenues. As users sample many gaming apps, often only spending minutes or days before disengaging, firms seek to entice as many new app adopters as possible to make an in-app purchase – making low-price approaches attractive. Indeed, anecdotal evidence and a survey among mobile gaming managers corroborate that low-price approaches are common. This study conjectures that this practice is not optimal despite strong conceptual arguments in its favor. Using bandit-based experimentation in a mobile game, the study develops its argumentation over several field experiments, ultimately showing that firms can achieve higher profit using a personalized skimming tactic. Findings indicate that personalization in freemium settings can be highly profitable, and that it has potential to increase engagement with content through increased access to premium experiences. From a public policy perspective, results highlight that the prohibition of low-price "gateway" offers may not be an effective measure to counter spending sprees in mobile games.

5.1 Introduction

Smartphones have become our constant and pervasive companions (Balasubraman et al. 2002; Einav et al. 2014; De Haan et al. 2018). They connect us to people, businesses, entertainment and media, they assist us in structuring and organizing our life, they help us pass time, lose weight, meditate and take medication on schedule. Most of this functionality is provided by mobile apps that can be downloaded on app stores such as Google's Playstore or Apple's Appstore (Arora et al. 2017). While many apps were historically only available against payment of a fee (Arora et al. 2017), this practice has been upended by the success of freemium pricing (Lambrecht and Misra 2016; Gu et al. 2018): The vast majority of apps can be downloaded and used for free and monetizes by means of premium upgrades offered through in-app purchases.

Facilitated by this pricing structure, smartphones' ubiquitous availability and apps' ability to spur the formation of strong habits (Block 2008; Pivetta et al. 2019; Jo et al. 2020), the "app economy" has become a mainstay of economic activity. Consumers downloaded apps 194 billion times in 2018 and spent $101 billion on in-app purchases in the same time period (App Annie 2018). Accounting for almost 80% of that revenue, gaming in particular has seen an unparalleled expansion of demand, additionally fueled by online social networks that facilitate viral sharing and network effects (Alsén et al. 2016; Sensortower 2019a). It is estimated that 50% of mobile app users play games regularly and that a global total of 2.4 billion people will play mobile games in 2019 (Kaplan 2019). This explosive growth has not only given rise to a large mobile gaming industry that is estimated to drive 60% of overall revenue in the gaming vertical in 2019 (App Annie 2018, p. 20), but to the prevalence of new types of consumer-firm interactions (Einav et al. 2014; De Haan et al. 2018; Tong et al. 2020). Many consumers only briefly sample a specific app and the average session length in mobile games is around ten minutes (median: six – Gameanalytics 2019).

To entice consumers to make a purchase during their mostly short interactions with a mobile game, many firms pursue low-price approaches. Most games offer their smallest in-app purchase for $0.99 and the median price point offered is $4.99 (McGregor 2015). Firms additionally use promotional offers – colloquially called "starter packs" or "beginner bundles," see Figure 5.1 – shortly after users download an app to increase the share of paying users. This practice seems reasonable as only a very small percentage of users makes any purchase at all and many consumers may initially be hesitant to spend money on virtual in-game goods, but literature offers no guidance to managers in how far their current practice is optimal or could be improved. This study draws on a longstanding stream of literature cautioning that low-price approaches can harm longer-term revenue generation (Lattin and Bucklin 1989; Blattberg et al. 1995; Mela et al. 1997; Dekimpe et al. 1998; Jedidi et al. 1999; Anderson and Simester 2004) and conjectures that current managerial practice may not be optimal. The analysis is developed over six large-scale field experiments in a popular mobile game and shows that pricing according to current managerial guidance indeed does not lead to highest profit.

Results and the used experimentation approach provide direct guidance to managers how they can more profitably price in-app purchase offers in mobile games. The experimentation approach also promises to be applicable well beyond the focal setting, e.g., to learn price paths when a new product is launched or for a subscription on a news website. The study further contributes to the literature by empirically studying price personalization (Pigou 2017; Rossi et al. 1996) in a highly connected online business-to-consumer (B2C) setting. To the authors' best knowledge, existing studies of the matter either work with model-based counterfactuals (Acquisti and Varian 2005; Shiller 2020) or are situated in a business-to-business (B2B) setting (Dubé et al. 2017a) where purchase decisions tend to be less emotional (Odlyzko 2004; Chatterjee and McGinnis 2010) and customer backlash less likely (Martinez 2014; Sinclair 2017). This paper further contributes to literature on

pricing by adapting skimming to the setting of mobile apps and evaluating its merits vis-a-vis a flat-price tactic in a fully randomized experiment (Shapiro 1983; Nair 2007). On the methodological front, the study speaks to recent advances in firm experimentation, specifically the use of bandit methods to aid in the efficiency of experimentation (Li et al. 2010; Schwartz et al. 2017; Sutton and Barto 2018; Misra et al. 2019). Reports on the application of such methods to marketing problems are rare but much needed. The current analysis is novel in concatenating several bandits with different reward specifications to emulate a pricing manager who attempts to implement a profitable skimming tactic.

The paper proceeds by providing conceptual background and background on the empirical setting, before presenting the analysis. To facilitate presentation, the analysis is structured into four studies that build on each other. Finally, the paper discusses results – including results of a survey among managers corroborating the widespread use of low-price approaches – and concludes.

5.2 Conceptual and Empirical Background

5.2.1 Pricing in-app purchases in mobile games

Similar to other freemium content, mobile games offer a basic version for free in perpetuity and premium upgrades such as virtual currency, boosts and other game enhancing virtual goods through in-app purchases and occasionally in subscriptions (Levitt et al. 2016; Lehdonvirta 2009; Runge et al. 2019). This particular flavor of freemium pricing is commonly called "free-to-play."

For the purposes of illustration, I will compare and liken the experience of a free-to-play gaming app with the experience at an amusement arcade. Consumers can enter the arcade (download the app) anytime, and freely select how long they stay. While arcades can charge an upfront fee, they often do not, similar to the

Figure 5.1: "Beginner's bundle" and in-app store in the mobile game Candy Crush Saga

Notes: New user promotion ("beginner's bundle") and in-app store selling premium virtual currency in the mobile game Candy Crush Saga.

download and start of most gaming apps. While the arcade will usually have opening and closing times and consumers will have to travel there, the app is available anytime and anywhere as people will commonly carry their mobile phone with them, virtually eliminating access frictions. Consumers will often visit the arcade in a group, but download the app by themselves – often at the recommendation of a friend or family member (Trusov et al. 2009; Alsén et al. 2016) – to then connect with other consumers (players) in the app if the app offers such functionality, and in online forums and chat apps auxiliary to the app.

Both the app and the store offer a number of complementary entertainment options, with gaming apps commonly offering a number of these completely for free in perpetuity – spurring continued sampling, the creation of habit and user retention (Bawa and Shoemaker 2004; Li et al. 2018; Eisingerich et al. 2019). Further (premium) entertainment options are available for a fee, through in-app purchases. Prices in the arcade will be uniform across consumers and possibly seasonally adjusted, e.g., to be lower in periods of low demand. Prices for (premium) experiences in the game app on the other hand can be controlled at the individual level as digital surfaces can be flexibly customized. The firm could charge each user a custom price. Further, while experiences in the arcade are a rival good and marginal cost of usage is greater than zero, in the gaming app they are non-rival and have zero marginal cost for the firm (Lambrecht et al. 2014). The price floors and supply mechanics in both settings are hence different: Any price greater than zero is profitable for a firm selling add-on experiences in a gaming app and it can sell infinitely much of an experience. Theoretically, in the gaming app, the firm could hence engage in first-degree price discrimination (Pigou 2017; Dubé and Misra 2017; Shiller 2020) and profitably charge each user her willingness to pay for a given in-app purchase as it does not incur marginal cost of production or distribution.

In reality, there are three main challenges to this approach: First, platform operators provide price tiers that are geographically "optimized" and binding to app publishers that hence cannot price continuously. Second, the firm does not know individual users' willingness-to-pay (Rothschild 1974). While platform operators observe plentiful information on users (location, demographic, behavior across the app ecosystem and mobile web, and even financial if offering a credit card such as Apple) app publishers can only observe some device and geolocation information and app use behavior of users in their app(s). This information allows a rough estimation of users' expected spending on in-app purchases and willingness-to-pay (Sifa et al. 2018), and allows for third-degree price discrimination, i.e., within customer

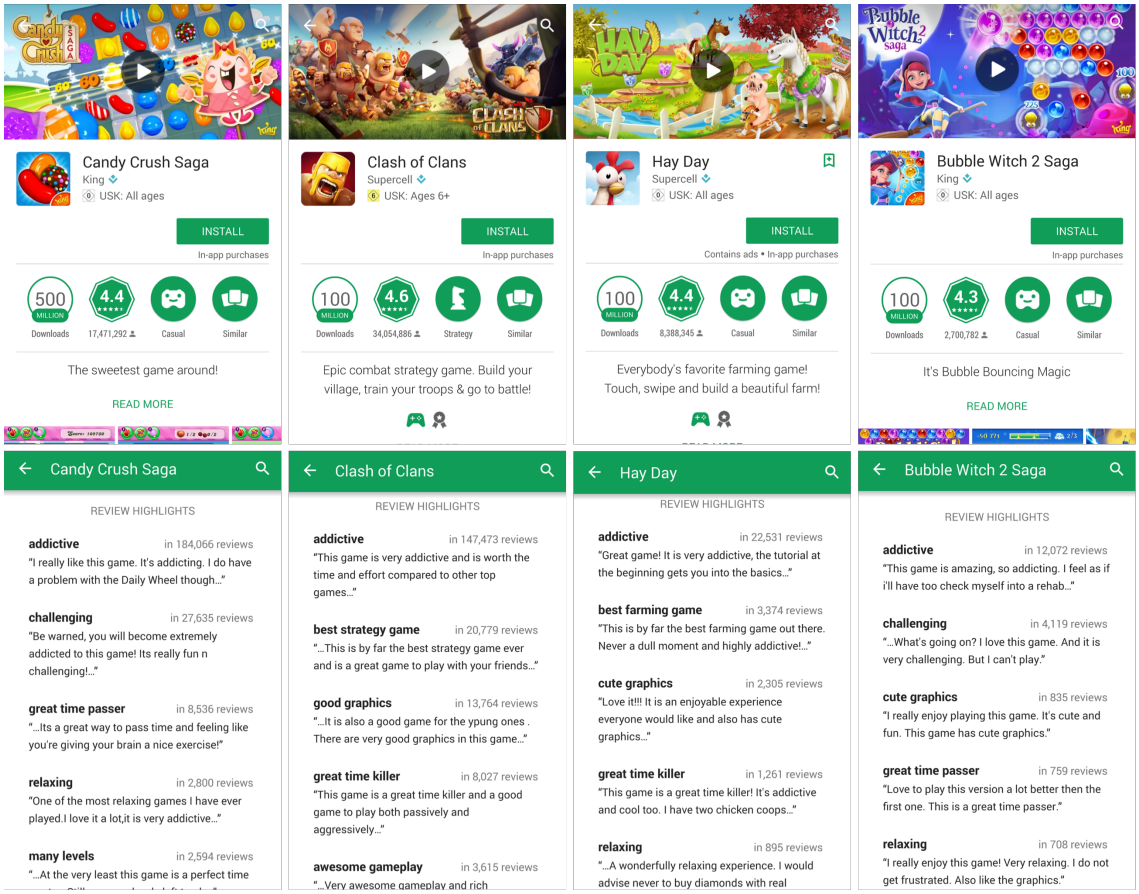
segments (Pigou 2017; Dubé and Misra 2017). Third, and most importantly: Mobile app users and especially gamers are very well-connected and share in-depth and ample information on forums and in chat apps (Cole and Griffiths 2007; Runge et al. 2019). Hence, there is a high degree of shared knowledge. Attempts to charge different prices for the same in-app purchase at the same time were historically quickly reversed by companies due to customer backlash (Martinez 2014; Sinclair 2017).

In practice, gaming apps usually offer different amounts of in-game virtual currency through in-app purchases at the price tiers set by the platform operator (most often Apple or Google). This in-game currency can then be used to buy in-game premium experiences such as boosts to beat levels, new characters or outfits, or other goods that enhance users' experience (Lehdonvirta 2009; Levitt et al. 2016; Gu et al. 2018; Li et al. 2018; Runge et al. 2019). Many publishers additionally offer promotional bundles of in-game goods and currency to users very early after app adoption (colloquially called "starter packs" or "beginner bundles"). These bundles are usually priced rather low (e.g., between $0.99 to $4.99) and promoted through in-app popups (e.g., see Figure 5.1) that often start appearing during first use of the app. They are akin to premium samples that intend to get users "hooked" on premium experiences and in-app purchasing (Bawa and Shoemaker 2004; Alter 2017; Eisingerich et al. 2019). The price of such a "starter pack" is the focal treatment manipulated in this study. As I vary price, I vary the quantities of bundled in-game goods to keep unit price fluctuations low, to ensure that users do not feel treated unfairly should they discuss the different versions of the offer (Huang et al. 2005; Chatterjee and McGinnis 2010; Li et al. 2019).

5.2.2 Research question

This study's main research question is: How can firms marketing mobile games use in-app purchase pricing and promotion to improve relevant economic outcomes, in particular monetization and engagement among new app adopters? While the scope

Figure 5.2: Word counts in customer reviews on Google's Playstore



Notes: "Addictive" is the by far most used word when consumers describe mobile game experiences.

of this analysis focuses on the perspective of firms, I extend the discussion to the perspectives of consumers and regulators in the final sections. In investigating this question, I focus on the evaluation of different pricing tactics for a new user in-app purchase promotion. Table 5.1 on page 121 gives an overview of the different versions of this promotion used in experimentation. I choose this approach because setting different (unit) prices for "normal" in-app purchases is practically infeasible as discussed in Section 5.2.1 on page 106. A main challenge is the high connectedness of users in this setting, making customer backlash very likely (Chatterjee and McGinnis 2010; Sinclair 2017). In the following, I will further develop the research question vis-a-vis existing literature.

5.2.2.1 Should the firm set a low or a high initial price?

Managerial practice seems to favor a low-price approach for starter offers as supported by a survey among 54 mobile game managers conducted by the authors: 25.9% of managers set a price lower than \$3, 59.3% a price lower than \$5 and 96.3% a price lower than \$10; excluding data scientists, these numbers are 27.5%, 65% and 100% of managers (N=40; question 10 on page 158). Managers believe they can sell users more after they have "hooked" them at an attractive lower price (Eyal 2014; Alter 2017). Risk aversion may be another reason leading managers to favor a low-price approach – I discuss the survey and its results in Section 5.5.2 and in Appendix A1 in more detail.

Four conceptual arguments support the choice of a low-price approach: First, online play is considered strongly habit forming (Eyal 2014). Several sources assert that mobile games spur strong habit formation and possibly addictive consumption patterns (Chen and Leung 2016; Kwon et al. 2016; Nevskaya and Albuquerque 2019; Jo et al. 2020). Figure 5.2 presents anecdotal evidence supporting this notion: "Addictive" and "addicting" are the by far most used words when consumers review popular mobile games. Kwon et al. (2016) analyze behavior in online social games using a rational addiction framework and find similar addiction coefficients as for alcohol or other substances. Jo et al. (2020) and Nevskaya and Albuquerque (2019) investigate how usage restriction policies impact consumer behavior in this setting. In this framing, a low-price approach for an initial "starter pack" seems optimal from the firm's perspective as it can serve as a gateway to more intense use and purchasing. The literature asserts that low-price approaches can sustainably increase demand for addictive goods (Becker and Murphy 1988; Becker et al. 1991; Katz and Lavack 2002; Chen et al. 2009; Gordon and Sun 2015). In fact, the practical viability of such an approach is the reason that some countries prohibit the sale of single cigarettes (Schütze 2014) and the retail sale of alcoholic beverages after certain hours (Hahn et al. 2010). Second, while freemium's zero-price point effectively attracts users

to adopt a product or service, here to download an app, it has been shown that consumers exhibit strong inertia around a zero-price (Shampanier et al. 2007; also termed "penny gap" Carter 2019). An attractive low-price offer is likely to be more effective in removing users from free use of the app and enticing them to spend money. Third, reduced time to process information and increased search cost in mobile apps favor a low-price approach. The attention a product can garner is an essential factor in consumers' decision to purchase it (Bettman 1979; Chandon et al. 2009). Sessions in gaming apps are short; Gameanalytics (2019) reports that users spend an average of ten minutes (median: six) in a mobile game before moving on to another activity or app (Yeykelis et al. 2014, 2018; De Haan et al. 2018). Additionally, mobile phone screens are small leading to increased search cost (Ghose et al. 2013). This combination of short attention spans and small surfaces reduces the ability to advertise in-app purchases, possibly making it harder to sell high-price items (Bemmaor and Mouchoux 1991). Fourth, a low-price approach can help alleviate uncertainty about quality. Virtual goods sold through in-app purchases are a new category that may be unfamiliar to customers (Hamari and Keronen 2017). A low-price approach can hence entice users to try out this new product category and reduce their uncertainty about it (Foubert and Gijsbrechts 2016).

Based on these arguments, current managerial practice seems highly appropriate. The firm should favor a low-price "penetration" approach to make in-app goods attractive to as many customers as possible during the short interaction time and small surface that are available, and to "hook" free users on premium experiences. Such an approach is likely to work particularly well if consumers struggle to self-regulate their consumption after an initial low-price purchase and then continue to purchase without too much regard to price. Indeed, a recent study on inter-temporal pricing in freemium games (Runge et al. 2019) suggests that a lack of consumption regulation may help explain how regular low-price offers lead to an expansion of primary demand in this setting.

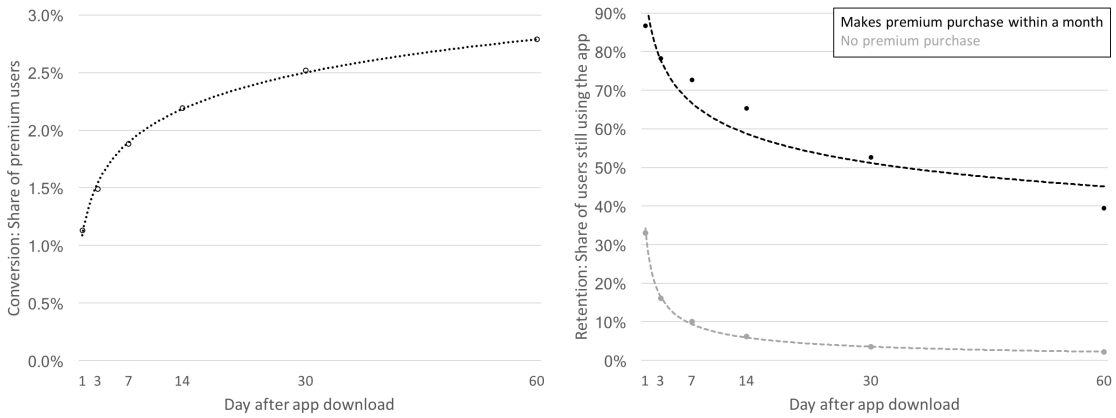
On the other hand, an extensive stream of literature cautions against the use of low-price approaches due to adverse mid- and longer-term effects (Lattin and Bucklin 1989; Blattberg et al. 1995; Mela et al. 1997; Dekimpe et al. 1998; Jedidi et al. 1999; Anderson and Simester 2004). A main argument pertains to consumer expectations: Lattin and Bucklin (1989) find that price promotions can dampen consumers' reference prices. Mela et al. (1997) find increased price sensitivity from price promotions for both loyal and non-loyal customers in the long run. Anderson and Simester (2004) find a negative long-run effect of price discounts on established customers. They point at forward buying, customer learning, and increased deal sensitivity as important longer term effects. In this framing, it could be argued that the firm should start app users off on large higher-price "starter packs" to ensure it does not adversely impact consumers' (price and quality) expectations and to form habits towards large premium purchases (Chen and Leung 2016). Regulators may prefer such an approach to protect consumers from "gateway" purchases if addictive potential of mobile games is high indeed (Schütze 2014).

Overall, there are arguments both in favor of a low- and a high-price tactic for a new user in-app purchase promotion from a firm's point-of-view, with managerial practice favoring the former. To assess the differential impact of a low-, mid- and high-price approach, Study 1 presents an analysis of an experiment randomly assigning 363,440 new app adopters to a \$2.99 (low), a \$4.99 (managerially preferred), a \$29.99 (high) in-app purchase promotion, or a control condition.

5.2.2.2 Can non-static pricing be of use to the firm?

So far, we have considered static price tactics, but the firm may opt to dynamically adapt price over time. E.g., skimming promises to combine a low- and high-price approach, by gradually lowering an initial higher price to sell the good to customers with increasingly lower valuations (Shapiro 1983; Acquisti and Varian 2005; Nair 2007). Skimming has potential to work particularly well when consumers are unin-

Figure 5.3: Conversion and retention in the studied app



Notes: Conversion, i.e., share of app adopters making an in-app purchase, and retention, i.e., share of app adopters still logging into the app, over days after app download; data are from the app used for analysis.

formed (Varian 1980) or myopic (Becker and Murphy 1988; Becker et al. 1991), i.e., when they do not actively seek out information on the market environment (from other consumers on online forums, chat apps, etc.) or when they do not pay much attention to what may happen in the future. It is unlikely that mobile game users are uninformed as discussed in Section 5.3.1.1, but they may act myopically (Kwon et al. 2016), possibly due to impaired consumption regulation (Kwon et al. 2016; Runge et al. 2019). Consumers with high valuations will then buy at a high price without regard to potential future price drops.

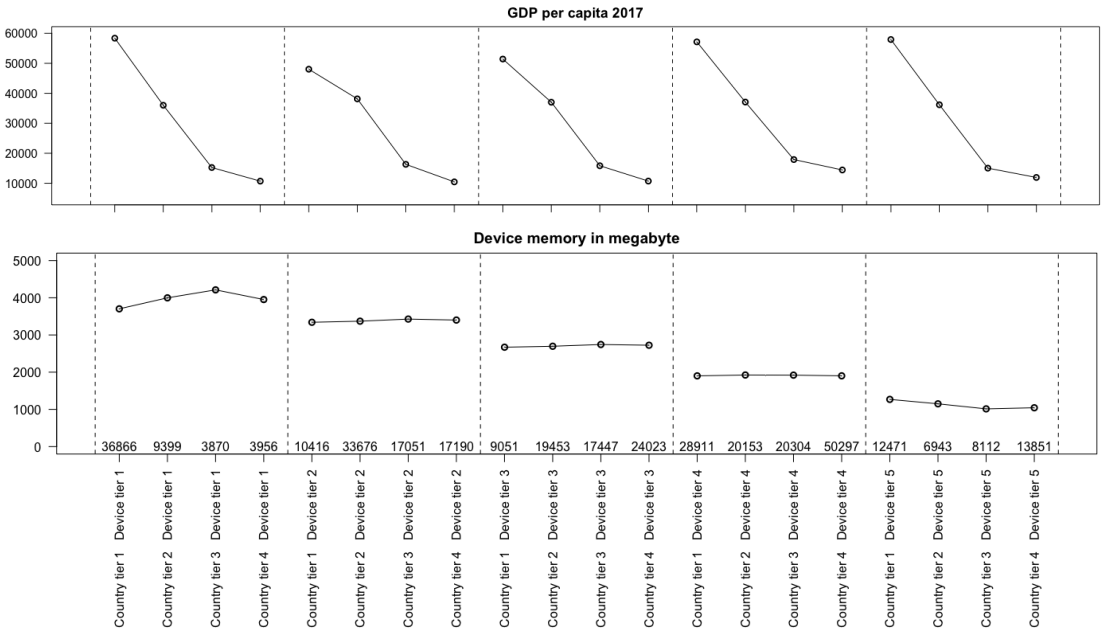
In this perspective, the firm will benefit from lowering the price and quantity of the offered bundle over time. Especially, as it has full agency over users' individual experiences (Ansari and Mela 2003; Arora et al. 2008), it can make a lower offer only to users who have not yet made a purchase in the app, after "locking in" users with higher valuations at a high initial price. The success of such a tactic is predicated on users with high expected spending (high valuation) making a purchase more quickly than users with lower valuation; and on users with low valuations for premium purchases retaining in the app beyond a few initial sessions, such that the firm can observe their unwillingness to purchase at a higher price and still make them an

offer at a lower price (which requires the users to still log into the app to see the offer as app notifications tend to be largely ignored; see Nevskaya and Albuquerque 2019).

Addressing the former requirement, we can look at average spending for users who convert on different days after app download: Users in the mobile game providing data for this study who make a first purchase on the first two days in the app spend $4.01 on average within a month after app download; the same figure is at $2.89 for users who make a first purchase between day three and seven after download, and at $2.28 for users who make a first purchase in their second week after app download.[1] Speaking to the second requirement, Figure 5.3 shows conversion and retention profiles for users in the mobile game studied here. As can be seen from the conversion profile, about 2.5% of app downloaders make a purchase in the game within 30 days after app download. The retention profile is shown separately for users who make a purchase within 30 days versus for those who do not. Unsurprisingly, purchasing users retain much more strongly with the app. In fact, only about five percent of initial downloaders who do not make a purchase within a month are still using the app two weeks after install. I use this insight in designing a simple skimming approach that initially offers the (managerially preferred) $4.99 bundle to users, a $2.99 bundle if they have not made a purchase a week after app download (when about ten percent of non-paying app adopters are still active), and a $0.99 bundle if they have not made a purchase two weeks after app download (when about five percent of non-payers are still active). I choose this approach to ensure that the lower-price offers still have potential to reach meaningful shares of non-paying users. To assess the merits of this skimming approach, study 2 randomizes 72,243 newly arriving users into a condition applying this simple skimming tactic and into a control condition receiving the $4.99 bundle in a flat-price approach.

[1]The reported figures are adjusted for days available to make purchases for users who convert later.

Figure 5.4: GDP per capita and device memory for country and device tiers



Notes: GDP per capita (2017, most recent available world development indicators when analysis was conducted (The World Bank 2018)) and device memory in mega byte across the device and country tiers used by the industry collaborator. The different tiers are effective in distinguishing device quality and price (as proxied with device memory) and levels of national wealth (as proxied with per capita GDP). 95% confidence intervals are not visible as they are too narrow. Segment size shown on lowest x-axis.

5.2.2.3 Can the firm use available data to personalize price?

In addition to lowering price over time for users who have not made a purchase, the firm may choose to set different prices for different users (Pigou 2017; Dubé and Misra 2017; Dubé et al. 2017a; Shiller 2020). The data available at app download are device (processor, memory) and geolocation information (aggregated at the country-level for privacy reasons); the data sponsor had generated device and country segments from these information for use in app marketing. Figure 5.4 depicts average device memory – a strong correlate of device price[2] and the strongest observed user-level correlate of expected user spending[3]– and average per capita gross

[2]For an overview of mobile phone technical specifications and price, e.g., refer to `https://www.phonearena.com/phones/compare`.

[3]A simple regression of revenue per user a month after app download on device memory as shown in model M4 in Table 5.4 on page 133 confirms this: An additional giga byte (1000 mega byte) of device memory results in almost an additional dollar of spend on premium in-app purchases a month after app download for a user on average.

domestic product (GDP) across user segments, substantiating that the segments capture heterogeneity in users' willingness-to-pay in highly complementary product categories (device to use apps) and in country-level income and economic development (Von Neumann and Morgenstern 2007; Du and Kamakura 2008). Building on these insights, the three panels in Figure 5.5 on page 119 show that users in different country and device tiers display very different in-app behavior, with premium demand being substantially higher in "higher" country and device tiers, both measured by the share of app adopters making a purchase and by the average amount spent on premium upgrades: While more than seven percent of app adopters make a purchase by day 30 after app download in the top country-device tier combination (N=36,866), the same number is close to zero in the lowest country-device tier combinations (N=20,304 and N=50,297 respectively). Many differences in conversion and revenue across segments are statistically significant as indicated by the 95% confidence intervals around the values one month after app download (top-most line in each panel). Differences in engagement (as measured by minutes spent in the app) are less severe, but still visible and partially statistically significant as shown in the bottom panel of Figure 5.5. It seems plausible that users in a segment with relatively higher expected spending will be more inclined to purchase a high-price promotion than users in a segment with lower expected spending and purchase value (Rossi et al. 1996; Acquisti and Varian 2005). Based on this proposition, it should be possible for the firm to profitably personalize the promotion offered to users in different segments. Study 3 takes this proposition to an empirical test by means of offline evaluation on the data generated in studies 1 and 2, and by means of bandit-based online learning. Study 4 builds on the findings of study 3 to propose a heuristic personalization and an experiment design to learn a personalized skimming tactic.

The viability of such personalized pricing of the promotional offer is further predicated on users either not informing themselves about the fact that other users

receive different promotions or users not feeling that the different promotional offers provide drastically different value for their money, i.e., that they are being treated unfairly (Chatterjee and McGinnis 2010; Li et al. 2019; also see free-form answers to the survey shown in Table 5.9 in Appendix A1 on page 157). As discussed in Section 5.2.1, (subsets of) users are likely to inform themselves and are well-connected. In collaboration with managers at the data sponsoring firm, the authors hence designed differently priced versions of the promotion such that they all offer an attractive discount to users compared to prices in the normal in-app purchase shop while keeping unit price differences small. Table 5.1 on page 121 shows the different versions of the promotion. Each bundle contains the premium in-game currency, another in-game currency and a few in-game goods. The content is awarded over a 14-day period during which the user has to log into the app to redeem her goods. This approach, combined with the attractive discount, intends to incentivize users to engage sustainably with the app and is widely adopted in the mobile game industry.

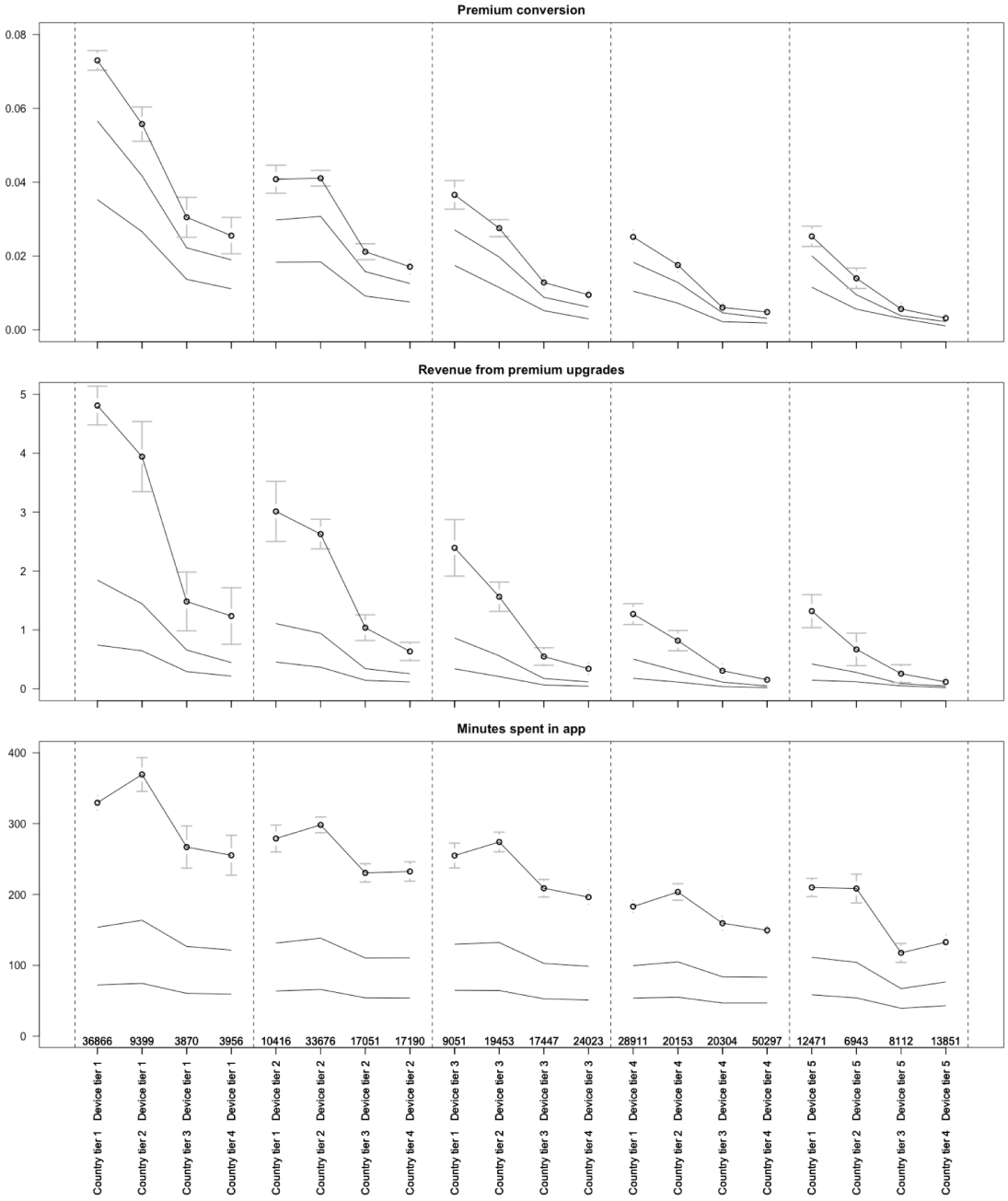
5.3 Method

5.3.1 Institutional details

5.3.1.1 In-app purchases in the game

The empirical setting for this study is a mobile gaming app that has been played by more than ten million users as of early 2019 and was part of the top grossing charts for all of 2018. Games as a category make up the absolute majority of app downloads worldwide (Sensortower 2019b, p. 27 and 28) and account for close to 80% of overall revenue from apps (App Annie 2018; Sensortower 2019a). The gaming app studied here is available on both the Google Playstore and the Apple Appstore. It uses the freemium/free-to-play model common in games and gamified apps such as, e.g., Tinder, that combines non-contractual (one-off) and contractual

Figure 5.5: Users' in-app demand by contextual segments observed at app download



Notes: In-app demand for different user segments as identified by meta data available at app download, showing behavioral slices after one day, seven days and 30 days after app download. Revenue in $ and premium conversion, i.e., share of users that made an in-app purchase, in percent. 95% confidence interval is only plotted for the 30-day slices to avoid clutter. Revenue has been winsorized at the 98th percentile among paying users to control for outliers. Segment size shown on lowest x-axis.

(subscription) in-app purchases. As common in freemium games, premium in-app purchases are offered in an in-app shop and comprise offers of in-game currencies that

can in turn be used to purchase in-game goods that facilitate and enhance gameplay (Lehdonvirta 2009; Levitt et al. 2016; Runge et al. 2019). The in-app shop offers bundles of 250 to 12,500 units of a premium in-game currency with prices ranging from $1.99 to $99.99. It further offers one bundle with the second in-game currency at $19.99. The base price for one unit of the premium in-game currency is 0.796 $-cents (based on the $1.99 offer in the app shop) and 0.0113 $-cents per unit of the second in-game currency. These values underlie the discount calculations in Table 5.1 on page 121.

5.3.1.2 The treatment

In addition to the in-app shop, the app offers premium upgrades through in-app promotional popups. The first and most frequent of these promotional popups that a user receives after app download offers a 14-day schedule of currency and in-game goods as outlined in Table 5.1. It comes in six variants with price points from $0.99 to $29.99. If a user purchases this offer, they are offered a step-up bundle at three times the initial price (hence with price points between $2.99 and $89.99) and with three times the initial content to spur repeat purchases. The ensemble of initial and step-up offer establishes the treatment used for personalization.

Most online games have strong communities that communicate on dedicated websites, chats, forums, and even in real-life meetups (Steinkuehler 2004; Cole and Griffiths 2007). The game studied here is no exception, making severe customer backlash a likely outcome of first degree price discrimination (Pigou 2017; Martinez 2014; Dubé and Misra 2017). Instead, I personalize the first and most important in-app offer to facilitate users' sampling of premium content and repeat purchasing (Bawa and Shoemaker 2004; Foubert and Gijsbrechts 2016). This approach enables the firm to serve different price points to different users while being fair in extending a large discount to all users and keeping unit price differences small. In collaboration with the firm, I monitored online forums and customer service communication while

Table 5.1: Versions of the promotion available for personalization

Premium currency	One-off price	Second currency	One-off price	In-game goods	One-off price	Overall one-off price	Offer price	Discount
100 per day	$11.14	2,000 per day	$3.16	Daily good	$1.74	$16.04	$0.99	93.8%
200 per day	$22.29	5,000 per day	$7.91	Same as above	$1.74	$31.94	$2.99	90.6%
200 per day	$22.29	10,000 per day	$15.82	Same as above	$1.74	$39.85	$4.99	87.5%
300 per day	$33.43	15,000 per day	$23.73	Same as above	$1.74	$58.9	$9.99	83.0%
400 per day	$44.58	20,000 per day	$31.64	Same as above	$1.74	$77.96	$19.99	74.4%
500 per day	$55.72	25,000 per day	$39.55	Same as above	$1.74	$97.01	$29.99	69.1%

Notes: Promotion content and comparison of offer price to the value of content if priced at baseline prices from the in-app store. Users have to log in for a consecutive 14 days to claim the content of the promotion, managers use this approach to incentivize users' regular engagement with the app. I disregard discount rates in calculation of the discounts, they are likely high in this setting (Kwon et al. 2016; Runge et al. 2019).

the personalization studies were active – they remained free from customer complaints and mentions of the offer. On a methodological level, this approach permits to assume single unit treatment value (SUTVA; Rubin 1978) which is necessary for the causal inference approach I use.

Finally, it should be noted that the offered discounts shown in Table 5.1 may seem very high. Readers should bear in mind that I calculate them based on unit prices of one-off purchases that make the whole purchased content available to the user immediately. The offer on the other hand requires the user to come back to the app for 14 consecutive days to claim the whole content. It hence uses the high discounts as an incentive to retain users with the app (Eisingerich et al. 2019; Appel et al. 2019). Other reports from this institutional setting suggest that this order of magnitude for discounts is common (Levitt et al. 2016; Runge et al. 2019).

5.3.2 The learning approach

The firm wants to identify the most profitable price policy for the in-app promotion out of a set of candidates. Firms commonly run fully randomized experiments (colloquially called A/B or split tests) with their digital content offering (Baker et al. 2014; Schwartz et al. 2017; Misra et al. 2019) that randomize users into different treatment conditions and can assist with this identification. The first two stud-

ies use this approach. A/B tests are costly however as consumers are exposed to non-optimal offerings for the whole duration of the test, especially when the test randomizes equally across the set of available prices or designs. Simple forms of reinforcement learning (Sutton and Barto 1999) such as multi-armed bandits can reduce the cost of experimentation by increasingly exposing users to the best performing variant being tested (Schwartz et al. 2017; Misra et al. 2019; Du et al. 2019). In their contextual form, bandits can choose the best performing variant per individual user context where the context is defined by observable user data (Bertsimas and Mersereau 2007; Li et al. 2010). In this study, I consider a contextual bandit approach to assist with policy learning. In case observed contextual features are not relevant to the association of treatment and reward, i.e., they are not relevant for personalization of the treatment if the firm wants to impact the given reward/outcome, this approach defaults to a simple bandit – in a way, A/B tests are "nested" within bandits (by setting the bandit's learning rate to full exploration) which are in turn "nested" in a contextual bandit.

I use the contextual bandit module of Vowpal Wabbit,[4] an open-source, fast, and large-scale machine learning library developed at Yahoo Research and acquired by Microsoft Research. I choose this implementation as it has been successfully applied in the field before (Li et al. 2010), is well documented in existing research (Dudík et al. 2011; Bietti et al. 2018), and is built for scale and production application.

5.3.2.1 Formalizing the decision problem

The firm's decision problem can be formalized as a stochastic (i.i.d.) batched bandit learning problem. I choose to use a simple epsilon-based learning strategy as it has been shown to perform well (Bietti et al. 2018; Du et al. 2019), can emulate an A/B test by setting epsilon to one, and to keep complexity low to allow for managerial control and buy-in. Each period t, the bandit decides on the allocation of share $1-\varepsilon$

[4]For additional information, `https://github.com/VowpalWabbit/vowpal_wabbit` – the library's Github page – offers extensive documentation.

of users arriving in this period to the different actions a_j (different offer variants), where users are defined by different contexts $x_i \in X$, per its policy π. The remaining share of users ε is allocated randomly across available actions a_j for exploration.

Importantly, and in distinction to many applied bandit problems (Li et al. 2010; Schwartz et al. 2017), the loss that each action produces in different contexts, $l(a_j, x_i)$ is not observed immediately but with a delay. A batch of users has time of period length $t - (t - 1)$ to make the decision to purchase the offer or not; these purchase decisions form the basis for the bandit's reward calculation. At the end of period $t + 1$, the bandit reinforces its allocation policy based on its decisions up to period t to minimize the loss resulting from its policy:

$$\arg\min \sum_{t=1}^{t=t} l_t(\pi(x_i)) \tag{5.1}$$

where $\prod$ is a set of policies $\pi : x_i \to a_j$. For every policy, it holds that $p_t(a_j) \in [0, 1]$. Further, as long as $\varepsilon > 0$, it holds that $p_t(a_j) > 0$, ensuring the availability of data for the bandit's policy learning.

The bandit then uses the policy resulting from this reinforcement learning step in period $t + 2$, in addition to the specified learning with rate ε, and the reinforcement loop is repeated. It should be noted that the problem is formulated in terms of loss minimization to follow Vowpal Wabbit's framing (Bietti et al. 2018). Alternatively, it could be formulated in terms of reward maximization by multiplying the cost/loss by -1. It should further be noted that, similar to the empirical approach in Dubé and Misra (2017), this approach is inherently Bayesian as the learner updates its beliefs about the best policy each period.

5.3.2.2 Policy estimation

I use an inverse propensity-weighted approach in policy estimation from randomized data. This estimator is used both to evaluate policy candidates on fully randomized

experimental (A/B test) data and by the bandit in estimating feedback across contexts from its random allocation of share of users ε to all available actions. Both estimations are counterfactual in nature and require an accurate model of the policy used for allocation (Dudík et al. 2011; Bietti et al. 2018) to obtain unbiased estimates using an inverse propensity-weighted estimators (Horvitz and Thompson 1952; Hitsch and Misra 2018):

$$\hat{l}_t(a) = \frac{l_t(a_j)}{p_t(a_j)} 1\{a = a_j\} \tag{5.2}$$

This estimator is unbiased for any $p_t(a_j) > 0$, but can have large variance when $p_t(a)$ is small. To lower said variance, Dudík et al. (2011) propose a combination of both approaches to make estimates more robust:

$$\hat{l}_t(a) = \frac{l_t(a_j) - \hat{l}(x_i, a_j)}{p_t(a_j)} 1\{a = a_j\} + \hat{l}(x_t, a) \tag{5.3}$$

This doubly robust estimator reduces variance if $\hat{l}_t(x_t, a_j)$ is a good estimate as the small numerator balances out a possibly small denominator. The second term ensures that the estimator is unbiased. I use such a doubly robust estimator as provided in Vowpal Wabbit (Dudík et al. 2011) for online bandit-based learning, and use the simpler estimator described in (2) (also see Hitsch and Misra 2018) for offline evaluation where I ensure that $p_t(a)$ is sufficiently large.

5.3.2.3 Priors relevant to learning

Ex ante, the firm does not know the demand function it faces (Rothschild 1974), but valuable prior information may still be available. The firm first needs to make the decision which offer to show to which user when users first start the app after downloading it from the app store. At this point, information as to users' geolocation and device characteristics are available as discussed in Section 5.2.1 (also see Sifa et al. 2018). Subsequently, the firm can observe how app users behave in the app,

i.e., if and how they make in-app purchases, play the game, interact with other players and consume advertising. This information can be used to update the initial decision. Subsequent decisions are however managerially constrained: The offer price can only be lowered or kept stable, not increased, to avoid upsetting customers. Further, while the fully randomized experiments in studies 1 and 2 start "cold" from a flat prior with equal exploration across treatments, reasonable managerial priors are used as a starting point for online learning. Such priors derive, e.g., from a product manager with expert knowledge on pricing strategies that can be used to "seed" the bandit. An example of such a prior is the practice to give a higher-price offer to a user segment that has historically seen higher spending per user as discussed in Section 5.2.2.3 and supported by other studies (Rossi et al. 1996; Acquisti and Varian 2005).

5.4 Studies

5.4.1 Study 1: Evaluating a high-, mid- and low-price tactic

Study 1 exposes app users to three different in-app offers and a control group reflecting the state of the app prior to the introduction of the offer. In the latter group, users received two one-off in-app promotional offers – one for the same premium in-game currency, and one for in-game goods (both priced at $2.99) – instead of the focal offer described in Table 5.1. A random 60% of 363,440 new adopters of the app during the first half of 2018 received the new in-app offer, being split equally across price points of $2.99, $4.99 and $29.99. The remaining 40% of users received the control treatment which reflects the state of the app prior to introduction of the offer.

In the analysis, I focus on intent-to-treat "slices" of user behavior as they accumulate until one day, one week and one month after initial app download. In

this intent-to-treat analysis, the denominator of the reported user behavior averages always is the number of users initially downloading the app (regardless if a user is still actively using the app), accounting for endogenous non-compliance in the calculation of averages. This approach is similar to Gordon et al. (2019) who present a more detailed exposition for the interested reader.

Table 5.2 on page 128 shows treatment main effects on user behavior outcomes such as demand for the focal promotional offer, wider demand for in-app premium content (both in $), advertising consumption and app use. Results show that a higher-price offer leads to lower conversion and a lower-price offer to higher conversion, in line with an inverse price-demand relationship. The two lower-price offers and the control condition all lead to a substantially higher offer and overall purchase incidence[5] (both well beyond 2% of the user base for the 30-day window) and also more repeat purchases than in the treatment condition with the high-price offer. This finding suggests that purchasing of premium content by more users leads to increased repeat purchases of such content, possibly due to uncertainty reduction (Foubert and Gijsbrechts 2016). It further confirms that the treatment has strong main effects and is hence a capable candidate for personalization.

It should further be noted that the two lower-price offer conditions ($2.99 and $4.99) that achieve higher free-to-pay conversion lead to statistically significant higher retention and time spent in the app than in the control condition. And, in a mean comparison, advertising consumption is highest in the condition with lowest conversion (the $29.99 offer condition). Advertising consumption is significantly higher in this condition than in the control condition which has significantly higher premium conversion, but equivalent app use. App use can be seen as independently driving advertising consumption: More time spent in the app means more time to be exposed to advertising. These results suggest that the new in-app offer is able to impact user engagement, possibly hand-in-hand with repeat purchasing;

[5] I use the terms premium purchase incidence and conversion (to premium) interchangeably. Both quantify what share of users or app adopters has purchased a premium upgrade.

and that higher premium conversion leads to reduced advertising consumption at similar levels of user engagement.

Finally, while the different treatments lead to substantially different outcomes in terms of users buying the offer, revenue generated by the offer and overall users making a premium purchase, effects on overall in-app purchase revenue are not significantly different between treatments. This result is sensible as the offer "only" accounts for between 4.2 to 24.1% of overall revenue (one month-result for the $2.99 and one day-result for the $29.99 treatment in Table 5.2) and variance of overall in-app purchase revenue is high as indicated by the large confidence intervals in Table 5.2.

5.4.2 Study 2: Evaluating a simple skimming tactic

From Study 1 (Section 5.4.1), we know that a high-price offer lowers premium conversion but increases per-user-revenue generated by the focal offer, and that a low-price offer increases conversion but reduces revenue per user from the offer. This section investigates if an increase in both offer conversion and offer revenue can be achieved with a simple skimming policy that lowers the initial offer price. To do so, 72,243 new app adopters all receive the $4.99 offer and then the price is dropped for a random 80% of users.[6] In this 80% treatment condition, users who have not made a purchase within the first week after app download, receive the $2.99 offer on day 8, and if they have not made a purchase within another week, they receive the $0.99 offer (as shown in Table 5.1) on day 14 after app download. Table 5.3 on page 130

[6]All experiments at the exception of this one were implemented in the app version published on the Google Playstore. There are two reasons for this approach: First, the app experienced much higher numbers of new downloads on the Google Playstore. Second, device information on Android devices (the ones accessing the Google Playstore) is much more granular and informative of user characteristics as a much higher variety of devices is available with many different price points. Apple, on the other hand, only releases a small number of new mobile devices once a year, substantially lowering the variety of mobile phones and price points available. On average, Apple devices have higher price points, suggesting that users with higher expected spending on premium upgrades should self-select into purchasing these devices. Indeed, Table 5.3 shows that average revenue on the Apple version of the app is much higher than on the Google version (see Table 5.2).

Table 5.2: Effect of differently priced offers on user behavior outcomes

Outcome window	One day after app download				Seven days after app download				30 days after app download			
Promotion price point	$2.99	$4.99	$29.99	Off	$2.99	$4.99	$29.99	Off	$2.99	$4.99	$29.99	Off
Offer conversion	1.28% (0.082%)	0.94% (0.07%)	0.22% (0.034%)	– –	2.0% (0.102%)	1.64% (0.092%)	0.37% (0.044%)	– –	2.63% (0.116%)	2.19% (0.106%)	0.6% (0.056%)	– –
Offer revenue (in USD)	0.013 (0.003)	0.047 (0.004)	0.066 (0.01)	– –	0.06 (0.003)	0.082 (0.005)	0.111 (0.013)	– –	0.079 (0.003)	0.109 (0.005)	0.179 (0.017)	– –
Overall paying users	1.41% (0.086%)	1.13% (0.077%)	0.67% (0.059%)	1.22% (0.056%)	2.22% (0.107%)	1.99% (0.101%)	1.28% (0.082%)	1.95% (0.071%)	2.93% (0.123%)	2.66% (0.117%)	1.91% (0.1%)	2.56% (0.081%)
Overall revenue	0.298 (0.069)	0.298 (0.074)	0.274 (0.04)	0.258 (0.028)	0.736 (0.127)	0.817 (0.156)	0.698 (0.094)	0.736 (0.073)	1.898 (0.288)	2.565 (0.508)	2.093 (0.302)	2.12 (0.214)
Overall revenue (winsorized)	0.21 (0.025)	0.211 (0.025)	0.231 (0.027)	0.218 (0.017)	0.511 (0.048)	0.546 (0.05)	0.546 (0.051)	0.561 (0.035)	1.365 (0.121)	1.528 (0.133)	1.502 (0.129)	1.484 (0.09)
Repeat purchases	0.009 (0.001)	0.008 (0.001)	0.006 (0.001)	0.008 (0.001)	0.025 (0.003)	0.027 (0.003)	0.02 (0.002)	0.03 (0.002)	0.07 (0.007)	0.082 (0.01)	0.066 (0.007)	0.081 (0.006)
Time spent in app (in minutes)	56.7 (0.694)	56.4 (0.685)	56.4 (0.676)	56.4 (0.48)	112.3 (2.015)	112.2 (2.01)	111.1 (1.969)	110.6 (1.38)	229.0 (6.43)	229.2 (6.379)	223.5 (6.22)	220 (4.306)
Game rounds played	9.45 (0.136)	9.43 (0.141)	9.43 (0.137)	9.41 (0.097)	18.58 (0.384)	18.62 (0.39)	18.47 (0.381)	18.36 (0.267)	32.25 (0.846)	32.31 (0.84)	31.83 (0.837)	31.37 (0.577)
Retention (% of users active)	34.4% (0.346%)	34.2% (0.345%)	34.3% (0.345%)	34.5% (0.244%)	11.8% (0.235%)	11.8% (0.235%)	11.6% (0.233%)	11.5% (0.164%)	4.87% (0.157%)	4.89% (0.157%)	4.66^ (0.153%)	4.65% (0.108%)
Ads viewed	0.725 (0.018)	0.726 (0.018)	0.743 (0.018)	0.731 (0.013)	0.946 (0.021)	0.947 (0.02)	0.966 (0.021)	0.943 (0.015)	1.051 (0.022)	1.047 (0.022)	1.062 (0.022)	1.045 (0.016)

Notes: Average outcomes per treatment group with 95% confidence interval in brackets. Overall revenue has a very skewed distribution due to the presence of extremely high-value users (Sifa et al. 2018), I hence also show revenue winsorized at the across-group 98th percentile of revenue among paying users. 72,407 users were allocated to the $2.99, 72,852 to the $4.99, and 72,671 to the $29.99 offer, 145,510 users remained in the control treatment that contained two separate one-off promotional offers in lieu of the in-app offer described in Section 5.3.1.2.

shows intent-to-treat effects of these policies on new app adopters seven, 14 and 30 days after app download. As can be gleaned from the first four result columns, the skimming policy achieves a lift in both offer conversion and revenue. The first result column in Table 5.5 on page 138 summarizes the percentage lift between the two treatment groups and applies a Bayesian significance test on this difference: Offer conversion significantly increases by 20.3% (prob(B>A) > 99.9%)[7] as well as offer revenue (+9.7%, prob(B>A) = 94.4%), with a significant positive effect on overall conversion (+9.8%, prob(B>A) = 98.7%). Effects on other relevant outcomes are directionally positive but not significant. While this simple skimming policy does not increase overall monetization (overall revenue remains unaffected), it has potential in increasing paying users with mild positive effects on revenue. Importantly, the firms' customer service and chat forums remained free from consumer comments or complaints that they would receive different offers over time, corroborating that skimming approaches have potential in this setting.

5.4.3 Study 3: Treatment effect heterogeneity and algorithm evaluation

5.4.3.1 Treatment effect heterogeneity

To investigate if the effects of differently priced offer variants tested in study 1 (Section 5.4.1) are heterogeneous across observed user characterists, I resort to simple linear models regressing longer-term (one month after app download) winsorized revenue on treatment indicators and the device information discussed in Section 5.2.2.3 and 5.2.2.1:

$$Y_i = b * T_{2.99,i} + c * T_{4.99,i} + d * T_{29.99,i} + e * DI_i + f * T_{2.99,i} * DI_i + g * T_{4.99,i} * DI_i + h * T_{29.99,i} * DI_i + a \tag{5.4}$$

[7] Prob(B>A) refers to the probability that the respective outcome is statistically significantly higher in the treatment group than in the control group, as obtained from a Bayesian significance test.

Table 5.3: Effect of skimming on user behavior outcomes

Outcome window	Seven days after app download		14 days after app download		30 days after app download	
Treatment group	Skimming (N=57,814)	Flat-price (N=14,429)	Skimming (N=57,814)	Flat-price (N=14,429)	Skimming (N=57,814)	Flat-price (N=14,429)
Offer conversion	3.04% (0.014%)	2.84% (0.027%)	3.54% (0.015%)	3.19% (0.029%)	4.17% (0.016%)	3.47% (0.030%)
Offer revenue (in $)	0.279 (0.016)	0.253 (0.030)	0.332 (0.018)	0.298 (0.033)	0.381 (0.019)	0.347 (0.037)
Overall paying users	3.57% (0.015%)	3.35% (0.029%)	4.18% (0.016%)	3.81% (0.031%)	4.98% (0.018%)	4.54% (0.034%)
Overall revenue	1.286 (0.181)	1.271 (0.440)	2.157 (0.307)	1.980 (0.647)	3.925 (0.626)	3.831 (1.234)
Overall revenue (winsorized)	1.057 (0.087)	0.981 (0.176)	1.662 (0.135)	1.490 (0.260)	2.870 (0.235)	2.700 (0.465)
Repeat purchases	0.045 (0.004)	0.042 (0.009)	0.075 (0.007)	0.066 (0.013)	0.139 (0.012)	0.125 (0.023)
Time spent in app (in minutes)	103.7 (2.075)	101.7 (4.098)	143.2 (3.321)	139.4 (6.518)	208.6 (5.879)	203.1 (11.44)
Game rounds played	16.54 (0.271)	16.17 (0.536)	20.93 (0.375)	20.41 (0.745)	26.88 (0.548)	26.26 (1.092)
Retention (% of users active)	12.6% (0.27%)	12.1% (0.53%)	9.0% (0.02%)	8.6% (0.05%	5.9% (0.02%)	6.1% (0.04%)
Ads viewed	0.894 (0.022)	0.860 (0.044)	0.948 (0.023)	0.911 (0.046)	0.993 (0.024)	0.973 (0.048)

Notes: Average outcomes per treatment group with 95% confidence interval in brackets. Overall revenue has a very skewed distribution due to the presence of extremely high-value users (Sifa et al. 2018), I hence also show revenue winsorized at the across-group 98th percentile of revenue among paying users.

where Y_i is user i's winsorized revenue until a month after app download, T_i is a binary indicator of the treatment the user was assigned to (the control condition is excluded as a reference category), and DI_i captures user i's device information, measured either as the firm's pre-defined device tiers or as device memory. I focus on device information as these present the strongest correlate of per-user expected value (Sifa et al. 2018), in particular device memory provides a continuous indicator of expected user value. I further opt for *98th percentile-winsorized* revenue as the outcome since a treatment priced between $2.99 and $29.99 is unlikely to meaningfully affect amounts of money spent that are in the thousands (a few users spend several thousand USD until a month after app download). A small and random difference in the number of these users allotted to randomized treatments can lead to spurious differences in means while overshadowing actual treatment-induced differences in the wider group of users. At the same time, it is important to include such high-value users in the analysis as one of the key prior beliefs pertains to a

high-price offer leading to better monetization outcomes among high-value users. Hence, to not exclude these users altogether, I impute the 98th percentile of revenue among paying users (approximately $342 for the one-month window) for the top 2% of paying users; at the lower end of paying users the winsorization does not change anything as as the 2nd percentile is virtually the same as observed values below it – the distribution of revenue per paying user is essentially flat at the lower end. The means and standard errors for revenue and winsorized revenue shown in Table 5.2 corroborate that this approach is able to reduce variance significantly.

Model M3.1 in Table 5.4 confirms the mean results shown in Table 5.2: The three different offers randomized in the A/B test do not result in significantly different revenue per user a month after app download. Further, model M3.2 substantiates insights from the visual analysis in Section 5.2.2.3: Users in higher-quality device segments spend significantly more on premium in-app purchases than users in lower-quality device segments. A month after app download, a user in the highest-quality segment "device tier 1" is expected to have spent $3.44 more than a user in the lowest device tier (which is excluded as a reference category) on average. Speaking to Section 5.2.2.3, Model M3.4 pinpoints device memory as the strongest exogenous continuous measure of expected premium demand of users: A user downloading the app to a device with 1,000 mega byte more memory is expected to spend $0.90 more on in-app purchases within the first month.

When it comes to identification of heterogeneity in treatment effects, Model M3.3 paints an interesting picture: Consistently across all segments, a higher-quality device segment interacts negatively with the low-price offer, but positively with the high-price offer, with the mid-price offer showing effects in both directions. These results suggest that a high-price (low-price) offer is comparatively better (worse) for generating revenue among high-value users. Results are, however, only partially and marginally significant. Model M3.5 provides a more direct test in using device memory instead of categorical device segments as an indicator for users' expected

spending on premium content. And indeed, the interaction of device memory with the $2.99 offer is statistically significant (see last column of Table 5.4). To be precise, a user on a device with 1,000 mega byte more memory will spend 20 $-cents less on premium content until a month after app adoption when placed in the low-price offer condition compared to in the control. This finding indicates that a low-price offer harms longer term revenue from users with higher expected value. Overall, these findings provide evidence confirming that a low-price offer has potential to decrease (increase) longer term revenue from high- (low-) valuation users. This finding also validates that a personalization of in-app offers has potential to be more profitable than a non-personalized offering.

5.4.3.2 Offline evaluation of the bandit learner

As prefaced in Section 5.3.2, I use Vowpal Wabbit's contextual bandit module. Before applying the learning algorithm in the field, I test its properties through offline evaluation similar to Dubé and Misra (2017) and Hitsch and Misra (2018). In actual field runs, I use an epsilon-greedy learning approach where the learning rate ε is set to simultaneously explore and exploit, and a batch is defined by 24-hour periods containing all users arriving in this time period. For the offline evaluation based on study 1's data, I apply an epsilon-first learning approach and define a batch by the sample available for learning.[8] Concretely, I use the first N=100,000 arrivals of study 1's experiment for learning (exploration) and the remaining N=117,930 for evaluation of the learned policy (exploitation) using 30 50% bootstraps. Device and country segments introduced in Section 5.2.2.3 constitute available contexts $x_i \in X$, and the randomized offers at $2.99, $4.99 and $29.99 establish available actions a_j.[9]

I evaluate two rewards (I transform them to the cost/loss that Vowpal Wab-

[8]Both approaches are applicable for a framing as a stochastic bandit problem which essentially assumes stationarity of the data-generating process over time. This assumption seems reasonable as all experimentation took place over a period of nine months during which the app experience remained fundamentally unchanged.

[9]Two managerially set boundaries should be noted here: 1) Firm-defined device and country segments are to be used as contextual data (and not, e.g., device memory and GDP per capita). The reason for this requirement is managerially requested consistency of the user experience and

Table 5.4: Regression of cumulative revenue a month after app download on treatment indicators and device memory

Dependent variable	Day 30 revenue (winsorized at 98th percentile)				
	M3.1	M3.2	M3.3	M3.4	M3.5
Treatment (Reference = Pre-offer state of the game)					
$2.99 offer	−0.112 (0.139)		0.065 (0.770)		0.0297 (0.156)
$4.99 offer	0.038 (0.616)		−0.051 (0.821)		−0.2149 (0.307)
$29.99 offer	0.021 (0.784)		−0.227 (0.310)		−0.2906 (0.167)
Device tiers (Reference = Device tier 5)					
Device tier 1		3.437 *** (0.000)	3.258 *** (0.000)		
Device tier 2		1.256 *** (0.000)	1.328 *** (0.000)		
Device tier 3		0.386 *** (0.000)	0.347 ** (0.033)		
Device tier 4		−0.034 (0.720)	−0.071 (0.633)		
Device memory					
(in mega byte)				0.0009 *** (0.000)	0.0009 *** (0.000)
Interactions					
Device tier 1 x $2.99 offer			−0.286 (0.335)		
Device tier 2 x $2.99 offer			−0.528 * (0.056)		
Device tier 3 x $2.99 offer			−0.100 (0.723)		
Device tier 4 x $2.99 offer			−0.013 (0.959)		
Device tier 1 x $4.99 offer			0.526 * (0.076)		
Device tier 2 x $4.99 offer			−0.002 (0.994)		
Device tier 3 x $4.99 offer			0.100 (0.722)		
Device tier 4 x $4.99 offer			−0.027 (0.917)		
Device tier 1 x $29.99 offer			0.657 ** (0.027)		
Device tier 2 x $29.99 offer			0.170 (0.539)		
Device tier 3 x $29.99 offer			0.193 (0.492)		
Device tier 4 x $29.99 offer			0.228 (0.379)		
Device memory x $2.99 offer					−0.0002 ** (0.036)
Device memory x $4.99 offer					0.0001 (0.200)
Device memory x $29.99 offer					0.0001 (0.115)
Intercept	1.438 *** (0.000)	0.583 *** (0.000)	0.625 *** (0.000)	−0.976 *** (0.000)	−0.936 *** (0.000)

Notes: Results of a linear regression of winsorized revenue a month after app download on treatment indicators and background data, plus interactions. The full sample of the historic random trial was used, hence N = 363,440. * significant at the 10%-level, ** 5%-level, *** 1%-level.

bit prefers as an input through multiplication by −1), offer revenue (r_1) and offer conversion (r_2)[10] and the resulting policies $\pi_{revenue}$ and $\pi_{conversion}$, using an inverse

offer prices within the segments used for wider marketing activities. 2) The no-offer condition is not available as a treatment for personalization due to high technical cost of simultaneously maintaining it with the offer conditions. I hence exclude it from offline evaluation, reducing the set of users from 365,440 to 217,930 users who were randomly allocated to one of the offers at a price of $2.99, $4.99 or $29.99 (see Table 5.1).

[10]Rewards are derived from user behavior on the day of app download and the day after, i.e., users have 36 hours on average to contribute to the calculation of the conversion and revenue reward. This reward formulation is delayed compared to rewards that are directly observed such as ad clicks (Li et al. 2010; Schwartz et al. 2017).

probability weighted profit estimator described in Section 5.3.2.2. To repeat the core intuition: In essence, the "policy overlap" between the random assignment from the experiment and the bandit policy's assignment is used, i.e., the pool of users where random and targeted policy make the same decision, to generate a counterfactual estimate of outcomes in a world where all users would have been assigned based on the policy learned by the bandit. Appendix A2 presents a detailed account of the implementation by showing the Python code used for offline analysis.

I conduct this offline evaluation to confirm that (1) the bandit is able to learn a policy different from random assignment and (2) the two different rewards lead to meaningfully different price policies, concretely the conversion-reward leads to a low-price and the revenue-reward to a high-price policy. Figure 5.6 summarizes key results. The left column shows results when offer revenue (r_1), the right column when offer conversion (r_2) is used as a reward. The conversion-reward leads to a low-price policy with most users (approximately 28k) assigned to the \$2.99 and approximately 22k users assigned to the \$4.99 offer. The bandit with revenue as a reward assigns about 14k less users to these lower price points and assigns most users (approximately 23k) to the \$29.99 offer that the conversion-bandit only assigns 9k users to. A bandit with offer conversion as a reward hence leads to an average price of about \$8 while a bandit with offer revenue as a reward leads to an average price that is almost twice as high. These results confirm that the chosen bandit algorithm (1) is able to learn a policy different from random assignment for both reward specifications, and (2) learns a high-price policy with revenue as a reward and a low-price policy when conversion is used as a reward. This confirmation is crucial to justify taking the bandit to online field runs. The next section describes results from an online pilot that I use to assess if the bandit is able to learn a personalization policy at app download, i.e., assigns users differently to price points based on their available contextual data.

Figure 5.6: Different rewards and resulting price policies

Notes: The upper panel shows price assignments performed by a bandit trained on data from Study 1, the lower panel shows policy-based lift in the respective reward over mean reward in the best unpersonalized (the $4.99 offer) treatment. The right column shows results when offer conversion (buy or not), the left column shows results when offer revenue (buy or not times the price point) is used as a reward. The former leads to a low-price policy, the latter to a high-price policy – as shown in the price assignments in the upper panel.

5.4.3.3 Online pilot

In close collaboration with the data sponsor's engineering team, the authors built an online reinforcement learning system with Vowpal Wabbit's contextual bandit module at its core that is able to make decisions based on country and device tier contexts in real-time at app download and can then update these decisions based on collected behavioral data during users' use of the app. Appendix A3 outlines how the system works in detail. To validate that this system performs in line with expectations and to confirm the findings from the offline evaluation reported in Section 5.4.3.2, I perform an exploratory online run of the contextual bandit for the price decision at app download with offer conversion as a reward. This setting reflects reward specification r_2 in Section 5.4.3.2, and hence I expect the bandit to

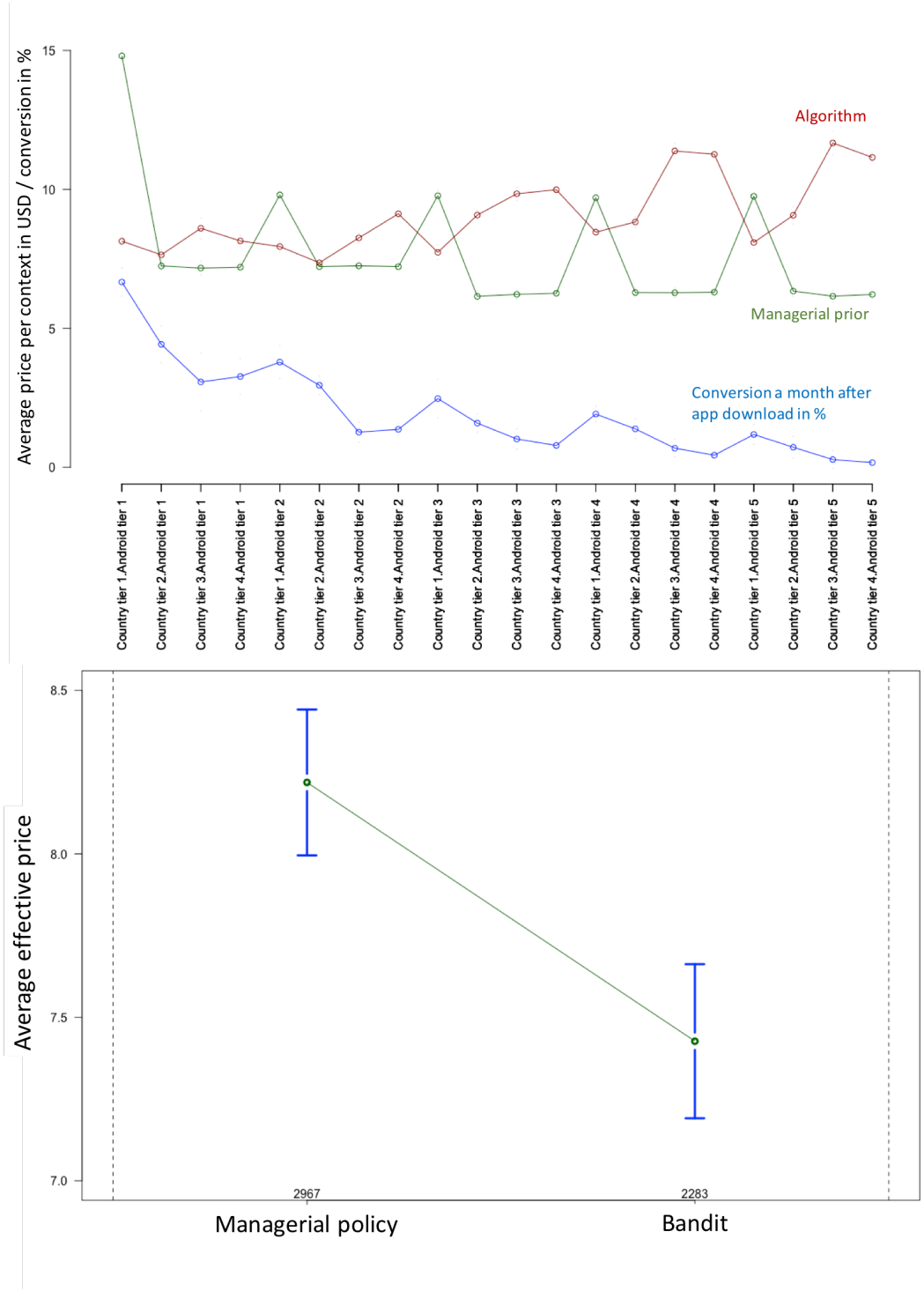
gear towards a low-price policy. While this approach can possibly harm longer term value generation from high-valuation users (per Table 5.4 and Section 5.4.3.1), it is expected to generate a larger number of paying customers than a high-price policy which was favored by managers, rather than the algorithm (possibly drastically) lowering the number of paying customers with a revenue reward.[11]

At app download, the device and country tiers used in marketing by the firm are available as contextual features as shown in Figure 5.5. In the offline evaluation in Section 5.4.3.2, I used 100,000 users to train the algorithm. To provide sufficient sample size to the bandit before starting its reinforcement loop, I seed it with a prior managerial policy (see Figure 5.7) and set the learning rate epsilon to 0.5, i.e., the bandit is allowed to randomly explore on 50% of arriving users. Once this policy has been executed on 200,000 arriving users (i.e., the bandit was able to explore on more than 100k users), I activate reinforcement. The bandit then updates its policy every day based on offer conversion within a day after app download as a reward, i.e., if a user made a purchase of the offer offered to her, either on the day of app download or the day after.

The offline evaluation results presented in Table 5.4 suggest that the bandit cannot learn a personalization policy based on the given contextual features as interaction effects between treatment and tiers are only marginally significant. And indeed the online reinforcement learning run confirms this expectation: Figure 5.7 shows that the algorithm implements a largely uniform price policy in the contexts with sufficient signal, and that the bandit seems to shift exploration with higher prices to contexts with limited or no signal, i.e., contexts with very low premium conversion among users. The right panel further shows that the bandit sells offers at a significantly lower price than the prior managerial policy and hence makes use of its agency in achieving a higher reward (=higher offer conversion).

[11] Creating consensus with and buy-in from managers proved crucial throughout the project as managers had to give up agency over price of the first and most important offer in the game – one of their key levers to exert managerial control over the game.

Figure 5.7: Price decisions per context by managers and algorithm



Notes: Average offer price charged by a managerial policy (plus 50% exploration) and by a bandit with offer conversion within a day after app download as reward. As expected, the bandit lowers the price compared to the managerial policy. Interestingly it appears to shift exploration with higher price points to lower country and device tiers. The reason for this behavior is the lack of signal in these contexts: The blue line shows that offer conversion (the relevant signal) is dismally low in these contexts, requiring more exploration with all available actions in these contexts.

Table 5.5: Effect of different price (personalization) policies on user behavior outcomes

	Study 2	*Study 3*		*Study 4*	
Outcome a month after app download	Simple skimming tactic (N=72,243)	Managerial prior with 50% exploration (N=242,604)	Conversion-bandit at app download, seeded with managerial prior (N=176,222)	Heuristic personalization at app download, price drops every two days with 50% exploration (N=104,052)	Heuristic personalization at app download and price drops every two days (N=100,821)
Offer conversion	+20.3% (>99.9%)	−19.1% (<0.01%)	−7.6% (3.0%)	+20.3% (>99.9%)	+44.7% (>99.9%)
Offer revenue	+9.7% (94.4%)	+1.5% (63.2%)	+3.2% (72.5%)	+4.2% (74.3%)	+15.9% (95.3%)
Overall paying users	+9.8% (98.7%)	−8.8% (0.2%)	+1.7% (66.4%)	+8.7% (97.5%)	+26.4% (>99.9%)
Overall revenue	+2.4% (55.2%)	−26.2% (3.1%)	+13.9% (77.8%)	+3.4% (58.7%)	+20.8% (81.7%)
Overall revenue (winsorized)	+6.3% (73.7%)	−8.5% (12.1%)	+12.2% (91.6%)	−0.2% (49.4%)	+23.2% (96.1%)
Repeat purchases	+11.4% (85.5%)	−9.2% (18.6%)	+10.9% (84.7%)	+8.5% (79.4%)	+12.9% (77.2%)
Time spent in app (in minutes)	+2.7% (80.2%)	−0.8% (34.0%)	+1.6% (76.5%)	+3.4% (87.9%)	+5.1% (93.9%)
Game rounds played	+2.3% (83.8%)	0.4% (60.1%)	+0.8% (69.4%)	+2.6% (89.4%)	+3.9% (95.6%)
Retention (% of users active)	−3% (20.2%)	−0.5% (42.0%)	+2.5% (80.8%)	+1.4% (64.3%)	+4.4% (84.3%)
Ads viewed	−2.7% (75.6%)	1.3% (79.8%)	−1.4% (19.9%)	+3.8% (93.9%)	+2.5% (81.9%)

Notes: Effect of different policies on user behavior outcomes by day 30 after app download as measured against a 20% randomized holdout group receiving the best-performing (in terms of revenue impact) unpersonalized offer at $4.99. Probability that the policy performs better than the holdout group as derived from Bayesian significance testing in brackets. The levels of the outcome variables per treatment group are shown in Table 5.6. Minor differences in percentage lift reported here versus as calculated on the values shown in Table 5.6 are due to rounding.

The second and third result columns of Table 5.5 show resulting user behavior outcomes a month after app download for the seeding, i.e., managerial prior with 50% exploration, and learning phase, i.e., active reinforcement learning based on conversion reward. To facilitate concise presentation and readability, results are shown as the relative percentage difference in user behavior compared to a randomized holdout group receiving the most profitable non-personalized offer for $4.99 (see "A/B test" results in Table 5.2).[12] The levels of the outcome variables per treat-

[12]The $4.99 offer generates highest mean revenue in the A/B test compared to the $2.99 and the $29.99 offers, and is hence most profitable due to zero marginal cost of the goods sold. Another A/B test run by managers compared it to the $9.99 and $19.99 offers where it also surfaced as causing the highest mean revenue. I do not use this A/B test in offline evaluation as the company did not track device and country information when this test ran.

ment group are reported in Table 5.6. Because the managerial prior policy charges relatively high prices compared to the randomized holdout group, the seeding phase leads to a statistically significant reduction in offer conversion by 19.1% (Prob(B<A) > 99.9%) and in overall paying users by 8.8% (prob(B<A) = 99.8%). Impact of the managerial policy combined with learning on repeat purchasing, overall revenue and hence profitability appears to be negative compared to the best flat-price tactic, with app usage and ad viewing remaining unchanged. During reinforcement learning, the conversion-bandit then lowers price on average (see Figure 5.7) reducing the offer conversion gap compared to the holdout group by a factor of 2.5, achieving a positive impact on profitability with marginal statistical significance (overall conversion, revenue, repeat purchasing). Impact on app use and ad viewing are not significant, but directionally in line with expectations (higher and lower respectively).

Overall, this online run confirms that the reinforcement learning system works as intended, that a bandit with conversion as reward implements a directionally profitable low-price policy, and that signal and treatment effect heterogeneity in the available contextual features are too low for the bandit to learn a personalized policy at app download.

5.4.4 Study 4: Personalized skimming

5.4.4.1 Devising a policy prior from managerial guidance and offline evaluation

As the bandit is unable to learn a personalization policy using available contextual features at app download, and offline evaluation on A/B test data is also underpowered to devise a personalization, I resort to guidance by the indications presented by offline evaluation: Higher price offers should be given to device and country tiers with higher expected spending which aligns with foundational marketing practice in purchase history- and geo location-based pricing (Rossi et al. 1996; Acquisti and

Table 5.6: Levels of user behavior outcomes in treated and control group

	Study 3				*Study 4*			
Outcomes a month after app download	Managerial prior with 50% exploration (N=242,604)		Bandit with conversion-reward at app download, seeded with managerial prior (N=176,222)		Heuristic personalization at app download, then price drops with 50% exploration (N=104,052)		Heuristic personalization at app download, then learned price paths (N=100,821)	
	Treated (N=169,605)	Control (N=72,999)	Treated (N=122,800)	Control (N=53,422)	Treated (N=82,728)	Control (N=21,324)	Treated (N=80,669)	Control (N=20,152)
Offer conversion								
	1.35%	1.67%	1.43%	1.54%	2.84%	2.36%	1.71%	1.18%
	(0.055%)	(0.093%)	(0.066%)	(0.105%)	(0.113%)	(0.204%)	(0.089%)	(0.149%)
Offer revenue								
	0.174	0.172	0.169	0.164	0.252	0.242	0.276	0.238
	(0.010)	(0.012)	(0.011)	(0.013)	(0.017)	(0.026)	(0.024)	(0.037)
Overall paying users								
	1.75%	1.92%	1.86%	1.83%	3.39%	3.12%	2.11%	1.67%
	(0.062%)	(0.100%)	(0.076%)	(0.114%)	(0.123%)	(0.233%)	(0.099%)	(0.177%)
Overall revenue								
	1.353	1.833	1.680	1.475	2.537	2.453	2.301	1.905
	(0.175)	(0.471)	(0.295)	(0.434)	(0.412)	(0.617)	(0.380)	(0.766)
Overall revenue (winsorized)								
	1.106	1.209	1.203	1.072	1.941	1.944	1.786	1.450
	(0.089)	(0.149)	(0.109)	(0.152)	(0.176)	(0.357)	(0.194)	(0.319)
Repeat purchases								
	0.051	0.056	0.057	0.052	0.095	0.087	0.121	0.105
	(0.007)	(0.009)	(0.006)	(0.009)	(0.009)	(0.016)	(0.011)	(0.029)
Time spent in app (in minutes)								
	142.9	144.1	161.6	159.1	210.5	203.6	165.8	157.8
	(2.942)	(4.607)	(3.805)	(5.626)	(5.464)	(10.27)	(4.752)	(9.035)
Game rounds played								
	17.02	16.96	19.07	18.92	23.55	22.95	19.51	18.77
	(0.252)	(0.387)	(0.319)	(0.481)	(0.438)	(0.830)	(0.386)	(0.749)
Retention (% of users active)								
	3.08%	3.09%	3.33%	3.25%	4.50%	4.44%	3.48%	3.33%
	(0.082%)	(0.126%)	(0.100%)	(0.150%)	(0.141%)	(0.276%)	(0.126%)	(0.248%)
Ads viewed								
	0.706	0.697	0.792	0.804	0.934	0.899	0.837	0.816
	(0.012)	(0.017)	(0.015)	(0.022)	(0.020)	(0.038)	(0.021)	(0.039)

Notes: Average outcomes per treatment group with 95% confidence interval in brackets.

Varian 2005; Du and Kamakura 2008). Managerial guidance further requested to not use the $29.99 offer (as it does not generate sufficient paying users) and to not use the very cheap $0.99 offer at app download. In agreement with managers, the authors hence devised a heuristic personalization policy that assigned offers ranging in price from $2.99 to $19.99 to user segments based on expected one-month revenue. The third column of Figure 5.9 on page 144 shows the resulting policy at app download.[13]

[13]The heuristic policy is derived from a mapping of price points to contexts based on their expected spend. The second panel of Figure 5.5 shows expectations for mean spend until a month after app download per context, as estimated from the A/B test data used in offline analysis. Es-

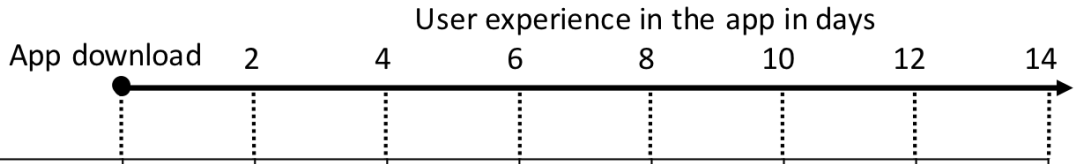
As already outlined in Section 5.3.2.3 managers further requested that offer price only be dropped and not increased to avoid customer backlash. As skimming proved to be profitable in Study 5.4.2, it seems promising to explore how we can best lower prices per user segment from the heuristically assigned prices at app download. To assist with learning per-segment price paths, I devise a learning system as shown in Figure 5.8 on page 142. I essentially concatenate contextual bandits to cover the entire period from app download to 14 days later. Each bandit receives two days of behavioral contextual data that are observed simultaneously with the reward of the previous bandit. Based on this behavioral data and the decisions of the previous bandit that are passed on, each bandit gets to decide to either drop the price or keep it stable for all users who have not yet made a purchase. We know from offline evaluation and the online pilot in Sections 5.4.3.2 and 5.4.3.3 that a conversion reward leads to a low-price and a revenue reward to a high-price policy. To emulate a skimming policy that initially sets rather high prices and then lowers prices, I set the first three bandits to have a revenue reward and the later three bandits to have a conversion reward. The idea is for the bandit system to first skim with high prices and then penetrate with low prices.

5.4.4.2 Evaluating the prior in an online run of the sequential bandit system

To take the system to an online run, I seed it with a prior to drop price at each decision stage and set the learning rate epsilon to 50%. At each decision stage, on a random 25% of users, price will hence be dropped and kept stable respectively (to account for 50% learning rate), and the bandit gets to make its own decision per user context (as defined by country tiers, device tiers and behavioral data) on the remaining 50% – with a prior to drop price. Choosing this "prior seeding" approach

sentially, contexts with an expected mean spend above $4 receive the $19.99 offer at app download, the $9.99 offer goes to contexts with an expected mean spend between $2.01 and $4, the $4.99 offer to contexts with an expected mean spend between $1.01 and $2.00, and the $2.99 offer to the remaining contexts with expected spend of $1 or below.

Figure 5.8: A schematic depiction of the firm's sequential decision problem



Decision point	App download	Beginning of day 2 of user experience	Beginning of day 4 of user experience	Beginning of day 6 of user experience	Beginning of day 8 of user experience	Beginning of day 10 of user experience	Beginning of day 12 of user experience
Actions	Subscription priced at: - 2.99 USD - 4.99 USD - 9.99 USD - 19.99 USD	- Keep price - Drop price	- Keep price - Drop price	- Keep price - Drop price	- Keep price - Drop price	- Keep price - Drop price	- Keep price - Drop price
Sample	All users	Users w/o purchase	Users w/o purchase	Users w/o purchase	Users w/o purchase	Users w/o purchase	Users w/o purchase
Contextual data available	- Geolocation - Device info	- Geolocation - Device info - Behavioral data on day 0 and 1	- Geolocation - Device info - Behavioral data on day 2 and 3	- Geolocation - Device info - Behavioral data on day 4 and 5	- Geolocation - Device info - Behavioral data on day 6 and 7	- Geolocation - Device info - Behavioral data on day 8 and 10	- Geolocation - Device info - Behavioral data on day 10 and 11
Reward	N/A	Purchase behavior on day 2 and 3	Purchase behavior on day 4 and 5	Purchase behavior on day 6 and 7	Purchase behavior on day 8 and 9	Purchase behavior on day 10 and 11	Purchase behavior on day 12 and 13
Price policy	Fixed policy based on historic spend per context	High-price; bandit has revenue-reward	High-price; bandit has revenue-reward	High-price; bandit has revenue-reward	Low-price; bandit has conversion-reward	Low-price; bandit has conversion-reward	Low-price; bandit has conversion-reward

Notes: Bandit-based system for online learning of price paths: At app download, a heuristic personalization policy is used and then six contextual bandits are concatenated with the same decision space (drop price or keep it stable), but different reward specifications to emulate a pricing managers that aims to implement a skimming policy. The three initial bandits have a revenue-reward to implement a high-price policy to skim, the following three bandits have a conversion-reward to implement a low-price policy to penetrate and convert users with lower willingness-to-pay.

avoids both the cold start problem (Li et al. 2010) and ensures that the action with highest expected profit – which is to drop price here to implement a skimming policy – is taken in lack of disconfirming evidence. The fourth result column in Table 5.5 shows results of this online run on user behavior outcomes as assessed on 104,052 new users who downloaded the app. While offer conversion and overall conversion are significantly up (by 20.3% and 8.7% respectively), other monetization and usage outcomes are not significantly impacted.

Importantly, the bandits do not take contextual information into account, but

Table 5.7: Regression of reward on sequential price decisions

Model	M5.1	M5.2	M5.3	M5.4	M5.5	M5.6
Decision point	Day 2	Day 4	Day 6	Day 8	Day 10	Day 12
Reward (= y)	Offer revenue	Offer revenue	Offer revenue	Offer conversion	Offer conversion	Offer conversion
Price decision						
Keep price (1)	−0.0027 (0.615)	0.0099 (0.046)	−0.0071 (0.099)	−0.0013 (0.085)	−0.0022 (0.007)	−0.0032 (0.015)
Intercept	0.0297 (0.000)	0.019 (0.000)	0.0173 (0.000)	0.0036 (0.000)	0.0032 (0.000)	0.0037 (0.000)
N	47,033	29,755	19,488	21,259	14,256	8,388

Notes: Results for a regression of the respective reward at each decision point on the price decision. Evidence does not support the prior to lower the offer price at the Day 2 and Day 4 decision points, but it does for the later decision points – leading to the skimming policy shown in Figure 5.9.

choose a uniform action across contexts. To verify that the bandits' decisions are reasonable, the authors conducted a more detailed analysis shown in Appendix A4. Here, I want to focus on informing a uniform – as contextual data do not matter to price paths in a consistent manner – skimming policy. To do so, I simply regress the respective reward at the six price decision points (see Figure 5.8) on an indicator if the offer price was lowered or not. Table 5.7 shows results; note that the sample size decreases for later decision points as the bandit only makes a decision for users that are active in the two days before the decision point and have not made a purchase yet. Models M5.1 and M5.2 indicate that there is no evidence supporting price drops at the Day 2 and Day 4 decision point. While the coefficient in model M5.1 is negative, it is far from significant. In model M5.2, the coefficient is positive – indicating that price should be kept stable. The coefficients in models M5.3 to M5.6 are consistently negative, indicating that price should be lowered at these decision points to achieve higher reward. These results hence support a price policy for the offer as shown in Figure 5.9. The initial price at app download is set based on the heuristic policy derived in 5.4.4.1 and then, at the first two decision points, offer price is kept stable and at the following decision points it is lowered.

Figure 5.9: The final personalization policy's price path per user context

Country tier	Device tier	App download	Day after app download					
			2	4	6	8	10	12
Country tier 1	Device tier 1	19.99	19.99	19.99	9.99	4.99	2.99	0.99
Country tier 2	Device tier 1	19.99	19.99	19.99	9.99	4.99	2.99	0.99
Country tier 3	Device tier 1	4.99	4.99	4.99	2.99	0.99	0.99	0.99
Country tier 4	Device tier 1	4.99	4.99	4.99	2.99	0.99	0.99	0.99
Country tier 1	Device tier 2	9.99	9.99	9.99	4.99	2.99	0.99	0.99
Country tier 2	Device tier 2	9.99	9.99	9.99	4.99	2.99	0.99	0.99
Country tier 3	Device tier 2	2.99	2.99	2.99	0.99	0.99	0.99	0.99
Country tier 4	Device tier 2	2.99	2.99	2.99	0.99	0.99	0.99	0.99
Country tier 1	Device tier 3	9.99	9.99	9.99	4.99	2.99	0.99	0.99
Country tier 2	Device tier 3	4.99	4.99	4.99	2.99	0.99	0.99	0.99
Country tier 3	Device tier 3	2.99	2.99	2.99	0.99	0.99	0.99	0.99
Country tier 4	Device tier 3	2.99	2.99	2.99	0.99	0.99	0.99	0.99
Country tier 1	Device tier 4	4.99	4.99	4.99	2.99	0.99	0.99	0.99
Country tier 2	Device tier 4	2.99	2.99	2.99	0.99	0.99	0.99	0.99
Country tier 3	Device tier 4	2.99	2.99	2.99	0.99	0.99	0.99	0.99
Country tier 4	Device tier 4	2.99	2.99	2.99	0.99	0.99	0.99	0.99
Country tier 1	Device tier 5	2.99	2.99	2.99	0.99	0.99	0.99	0.99
Country tier 2	Device tier 5	2.99	2.99	2.99	0.99	0.99	0.99	0.99
Country tier 3	Device tier 5	2.99	2.99	2.99	0.99	0.99	0.99	0.99
Country tier 4	Device tier 5	2.99	2.99	2.99	0.99	0.99	0.99	0.99

Notes: Offer prices of final personalized skimming policy, per country and device segments and decision point along users' experience in the app. The offer price shown here is the one made available to users who have not yet made a purchase and log into the app. The corresponding offers are shown in Table 5.1 on page 121.

5.4.4.3 Evaluation of the personalized skimming policy

The impact of the price personalization policy derived in the previous two sections on user behavior outcomes is shown in the last column of Table 5.5, as evaluated against the best flat-price tactic in a field run with 100,821 new users downloading the app. I find statistically significant effects (>95% probability that the treatment value is greater than the value in the non-personalized holdout group in a Bayesian significance test with flat priors) on offer conversion (+44.4%, prob(B>A) > 99.9%), offer revenue (+15.9%, prob(B>A) = 95.3%), overall paying users (+26.4%, prob(B>A) > 99.9%), and game rounds played (+3.9%, prob(B>A) = 95.6%). Impact on overall revenue is positive (+20.8%, prob(B>A) = 81.7%), but only significant for 98th percentile winsorized revenue (+23.2%, prob(B>A) = 96.1%). Other measures of app engagement show positive directional effects: +4.4.%, prob(B>A) = 84.3%, for

Table 5.8: Regression of monetization outcomes a month after app download on policy indicator and continuous background characteristics

Model	M6.1	M6.2	M6.3	M6.4	M6.5
Dependent variable	Revenue	Revenue (winsorized at 98th percentile)	Conversion	Offer revenue	Offer conversion
Treatment (Reference = Flat-price holdout)					
Personalized skimming	−0.416 (0.355)	−0.471 (0.033)	0.002 (0.384)	−0.047 (0.086)	0.003 (0.133)
Purchase history / expected user spending based on device data					
Device memory in giga byte	0.493 *** (0.000)	0.324 *** (0.000)	0.008 *** (0.000)	0.057 *** (0.000)	0.005 *** (0.000)
Interactions					
Personalized skimming x device memory	0.195 (0.214)	0.217 *** (0.005)	0.000 (0.669)	0.021 ** (0.031)	0.000 (0.549)
Intercept	−0.260 (0.502)	−0.073 (0.703)	−0.002 (0.236)	−0.021 (0.377)	−0.002 (0.332)

Model	M6.6	M6.7	M6.8	M6.9	M6.10
Dependent variable	Revenue	Revenue (winsorized at 98th percentile)	Conversion	Offer revenue	Offer conversion
Treatment (Reference = Flat-price holdout)					
Personalized skimming	−0.376 (0.290)	−0.091 (0.604)	0.002 (0.175)	−0.011 (0.620)	0.005 *** (0.004)
Country-based budget constraint / expected user spending based on geolocation data					
GDP per capita 2017 (in $10k)	0.325 *** (0.001)	0.298 *** (0.000)	0.005 *** (0.000)	0.044 *** (0.000)	0.004 *** (0.000)
Interactions					
Personalized skimming x GDP per capita 2017	0.213 ** (0.048)	0.094 * (0.077)	0.001 (0.225)	0.011 (0.106)	0.000 (0.726)
Intercept	0.109 (0.730)	−0.050 (0.751)	−0.003 * (0.093)	0.005 (0.789)	0.001 (0.461)

Notes: Results of linear regressions of monetization outcomes a month after app download on policy indicator, continuous background characteristics and interaction terms; N = 100,821; * significant at the 10%-level, ** 5%-level, *** 1%-level.

retention, and +5.1%, prob(B>A) = 93.9%, for time spent in the app. The personalized skimming approach is further able to mitigate any possible cannibalizing effect of increased premium purchasing on the consumption of ads (ads viewed are not impacted with +2.5% at prob(B>A) = 81.9%). Overall, this paints a very positive picture for the effects of the identified personalization policy on app monetization and engagement. As marginal cost of additional in-app purchases are zero for the firm, the policy is profitable compared to a counterfactual where the firm offers the most profitable non-personalized in-app offer at $4.99.

5.4.4.4 Policy effects by country and device characteristics

To understand the relative effectiveness of the personalized skimming policy (as shown in Figure 5.9) among users from different contexts, Figure 5.10 depicts mean lift in monetization outcomes in the policy treatment compared to the randomized $4.99 flat-price holdout group. As sample size is small within individual contexts, the figure shows outcomes for country and device tiers separately. It highlights how the skimming strategy drives down offer and overall purchases in high value device and country tiers early after app download. As prices then drop, this gap closes and ultimately results in a lift in purchases a month after app download. The strategy further leads to a consistent lift in offer and overall revenue in high-value segments compared to a flat-price strategy.

In lower-valuation segments (country tier 3 and 4, device tier 3 to 5), the lower relative initial price ($2.99 versus $4.99 in the holdout) leads to a strong increase in offer and overall purchases. This strong premium conversion lift is accompanied by a decrease in relative revenue in the lowest country and device tier; this revenue gap however closes over time after app download as purchasing lift remains strong and consistent. A month after app download, the personalized skimming policy has achieved a consistent high lift in the number of paying customers and mean revenue compared to the best-flat-price strategy applied in the holdout group.

Table 5.8 presents a multivariate assessment of the policy's effect on app monetization. Regressions of key monetization outcomes a month after app download on a treatment indicator and the continuous background variables device memory and GDP per capita confirm the strong main effect of device quality and GDP on consumers' in-app purchasing. The positive interaction effects in models M6.2, M6.4, M6.6 and M6.7 indicate that the policy's effectiveness is in large part driven by high-value users on higher quality devices and from higher-GDP countries.

Figure 5.10: The policy's effect on purchase behavior by segment

Tier	N	Device memory	GDP p. capita 2017	Offer purchases					Overall purchases					Offer revenue					Overall revenue				
				Day after app download					Day after app download					Day after app download					Day after app download				
				1	3	7	14	30	1	3	7	14	30	1	3	7	14	30	1	3	7	14	30
Country tier 1	19,822	2,821	57,165	-27%	-6%	9%	20%	26%	-14%	1%	20%	11%	8%	37%	59%	53%	47%	39%	3%	20%	41%	28%	16%
Country tier 2	15,339	3,081	38,118	1%	4%	-3%	13%	20%	17%	-6%	-5%	5%	6%	20%	15%	-1%	2%	-3%	74%	-10%	12%	25%	22%
Country tier 3	33,800	2,635	11,868	79%	93%	98%	149%	126%	39%	53%	46%	48%	42%	15%	23%	24%	29%	9%	-26%	-1%	-1%	-12%	16%
Country tier 4	31,860	2,362	16,749	10%	21%	3%	30%	48%	-20%	7%	12%	29%	46%	-28%	-22%	-34%	-30%	-28%	-54%	-26%	-9%	7%	50%
Device tier 1	12,176	4,734	40,359	-41%	-15%	5%	18%	16%	-30%	-12%	11%	22%	35%	53%	86%	77%	76%	55%	3%	18%	44%	52%	60%
Device tier 2	22,004	3,491	25,368	-13%	-7%	-1%	16%	29%	-1%	3%	4%	-2%	-10%	0%	4%	-1%	1%	0%	18%	-4%	6%	-7%	-24%
Device tier 3	21,075	2,769	20,143	65%	91%	40%	76%	75%	43%	52%	30%	20%	23%	8%	24%	-10%	-6%	-13%	-5%	-14%	12%	5%	54%
Device tier 4	31,056	1,855	26,566	73%	44%	33%	50%	62%	31%	7%	16%	15%	6%	22%	-1%	-9%	-14%	-16%	-1%	-9%	5%	16%	8%
Device tier 5	14,510	1,066	24,695	1%	39%	56%	105%	96%	5%	38%	48%	74%	64%	-39%	-17%	-7%	5%	-7%	-22%	34%	68%	68%	55%

Notes: Effect of the personalized skimming policy shown in Figure 5.9 on page 144, by device and country segments. The value shown is the lift in % in the respective outcome in the personalized over the flat-price holdout condition.

5.5 Discussion

The present study investigates a problem currently faced by mobile marketers: How can firms marketing mobile games use in-app purchase pricing and promotion to improve relevant economic outcomes, in particular monetization and engagement among new app adopters? Due to fairness considerations and the risk of customer backlash, the study does not alter the prices charged to different users directly, but devises differently priced "beginner bundles" that are targeted to new app adopters. It develops an analysis of price setting for such in-app promotions in mobile games over six large-scale field experiments showing that low price points can dampen monetization of users with high expected spending (study 1 in Section 5.4.1), that skimming has potential in this setting (study 2 in Section 5.4.2) and that price points higher than the ones considered by current managerial practice have merit among select users (study 4 in Section 5.4.4). Ultimately, the authors devise a heuristically personalized skimming tactic that achieves a 44.4% increase in demand for the "starter pack" (prob(B>A) $>$ 99.9%) and a 15.9% increase in revenue generated from it (prob(B>A) = 95.3%) compared to a randomized control group that receives the best performing non-personalized pack. This final personalization policy further achieves a 26.4% lift in overall paying users (prob(B>A) $>$ 99.9%), a 23.2% increase in winsorized revenue (to control for outliers; prob(B>A) = 96.1%) and 3.9% more

game rounds played (prob(B>A) = 95.6%) compared to the best flat-price tactic.

The authors want to use the discussion to highlight contributions to the literature and further develop the following aspects of the study before addressing limitations of this research: (1) 96.3% (100% if data scientists are excluded) of the 54 surveyed managers set a price below $10 for an in-app purchase promotion to new adopters ("starter pack" or "beginner's bundle") – the study shows that prices above $10 can be profitably charged to select user segments. This finding begs the question why managers seem to have such a strong "low-price bias." (2) The proposed method supports learning of a profitable personalization policy, but it ignores contextual data (Li et al. 2010; Bietti et al. 2018) – warranting further discussion. (3) Results show how personalization can increase firm profits, but what about implications from a wider societal perspective?

5.5.1 Contributions to the literature

The present study is one of the first to empirically investigate the feasibility of price personalization (Rossi et al. 1996) in an online business-to-consumer (B2C) setting. Empirical applications in online B2C settings are rare due to the high connectedness of users and related high risk of customer backlash (Chatterjee and McGinnis 2010; Martinez 2014; Sinclair 2017). Dubé and Misra (2017) is a rare exception, but their study is situated in a business-to-business setting (B2B), where implementations of personalized prices tend to be more accepted, e.g., due to lower emotionality of the purchase process and weaker fairness considerations (Odlyzko 2004; Chatterjee and McGinnis 2010). Similarly, Acquisti and Varian (2005) and Shiller (2020) study price discrimination in a B2C setting, but rely on model-based counterfactuals rather than field experimentation. Speaking to this literature, the present study shows that personalized promotion of in-app purchases (similar to coupon targeting in Dubé et al. 2017a) can increase user engagement, overall realized premium demand and app profitability, while not causing customer complaints and backlash. These

findings suggest that price personalization in online B2C settings more generally and in freemium settings more specifically can be highly profitable, and that it has potential to increase engagement with content through increased access to premium experiences. It however appears of paramount importance that firms take fairness considerations seriously and do not charge different prices to different users at the same time without good reason (Martinez 2014; Sinclair 2017). In digital settings, reasons such as different cost of supply do commonly not apply as marginal cost of production and distribution tend to be virtually zero. Extending on the study's contribution to literature on firms' price setting, to the authors' best knowledge, the present study is the first to evaluate a skimming approach in the field by means of actual randomization versus a flat-price control condition (Shapiro 1983; Nair 2007).

On the methodological front, the study speaks to recent advances in approaches to field experimentation, specifically the use of bandit methods to lower the cost of experimentation (Li et al. 2010; Schwartz et al. 2017; Sutton and Barto 2018; Misra et al. 2019). Reports on the application of such methods to marketing problems are rare but much needed. Furthermore, to the authors' best knowledge, the current analysis is the first to concatenate several bandits with different reward specifications to mirror a pricing manager who attempts to implement a profitable skimming policy. The suggested experimental design is applicable more widely, e.g., when a firm wants to learn profitable price paths for the launch of a new durable product or for a subscriptions on a news website.

Finally, the present study contributes evidence on the viability of price as a policy measure to counter excessive consumption of mobile online content, complementing a burgeoning literature on usage restrictions in this setting (Kwon et al. 2016; Nevskaya and Albuquerque 2019; Jo et al. 2020). Findings suggest that price and the prohibition of "gateway" offers (Schütze 2014; Wang et al. 2016) are likely not effective policy measures and that restrictions of usage and design elements such as lottery-based rewards may be more promising levers towards curbing unhealthy

consumption patterns (Koeder and Tanaka 2017; Nevskaya and Albuquerque 2019; Jo et al. 2020). Section 5.5.4 will further address this perspective.

5.5.2 Why do managers have a "low-price bias?"

A survey that the authors conducted in 2019[14] found that 96.3% of managers believe that extending a promotion to app adopters is essential to drive monetization in this setting – which is in line with recent research on the topic (Runge et al. 2019). 85.2% of managers further "focus on making an attractive offer to get users to make a purchase quickly," while only 14.8% think they should "focus on making users an offer at the highest possible price they are willing to pay" (question 7 on page 158). This managerial practice may originate from an overconfident belief in the ability to retain customers once they have made a purchase: The majority of respondents (61.1%) believe that they can sell users more later on and are not worried about potential adverse effects on consumer expectations (question 8 on page 158).

A further, and possibly most powerful, driver of the documented low-price bias is risk aversion: A low-price approach leads to much larger numbers of paying users but barely impacts overall spending of heavy users (see Section 5.4.1). It can hence appear to be less risky as revenue generation seems to be more evenly spread among users when only considering the mean and ignoring further moments of the distribution. To check if this mechanism may underlie the observed low-price bias, the authors included the following question in the survey (question 9 on page 158): "If you had the choice, which monetization configuration would you prefer? (1) Average lifetime value of users in the app is 10 USD; 1% of users spend 1000 USD each. (2) Average lifetime value of users in the app is 10 USD; 10% of users spend 100 USD each. (3) I'm indifferent between both options." 81.5% of managers chose option 2, 13% chose option 3, and only 5.5% chose option 1 (the order of response options

[14]Please refer to Appendix A1 for details on the survey's implementation and to Table 5.9 on page 157 for survey responses.

was randomized). Managers essentially can self-service their preference by choosing a low-price approach that lowers average spend per payer, but creates many more paying users as shown in Section 5.4.1.

As outlined in Section 5.2.2, the survey surfaced that 25.9% of managers set a price smaller than \$3, 59.3% a price smaller than \$5 and 96.3% a price smaller than \$10 for an initial promotion in a mobile game (question 10 on page 158). Excluding the 14 surveyed data scientists, these numbers are 27.5%, 65% and 100% of managers. Within current managerial practice, the presented personalization policy would hence be impossible to devise for lack of price points greater than or equal to \$10. Further, only about half of surveyed managers (53.7%) have used any sort of price personalization (question 14 on page 159). It hence appears that the presented personalization approach can have major impact on managerial practice and provide direct guidance how to generate higher profits in this setting.

5.5.3 Effectiveness of learning algorithm

The authors chose a contextual bandit learner as it had been documented to be effective in previous literature, both as a wider class of algorithm applied to marketing problems (Schwartz et al. 2017; Misra et al. 2019) as well as the specific implementation in Vowpal Wabbit (Li et al. 2010; Bietti et al. 2018). It was able to assist with learning a profitable personalization policy, but ultimately defaulted to the non-contextual version of a bandit in that available contextual data were not effective towards learning of a personalized policy. This result is disappointing, but does not take away from the essential premise that the presented learning system can learn profitable skimming policies well beyond the setting studied here. To the authors' best knowledge, the concatentation of several bandits that pass their decision on to their successor and have different rewards is novel and akin to building an "artificial pricing manager" who automatically adjusts offer price based on observed demand characteristics (Rothschild 1974). When devising such a system,

a few aspects should be considered: Bandits should be cleanly delineated in a new bandit's agency only beginning once the reward of the previous bandit has been fully observed. Reward observation windows should further be sensible via-a-vis product usage and purchase cycles. Rewards should also be chosen to mimic a substantively sensible policy maker; in this study: a policy maker that skims high valuations by aiming at high revenue generation and then penetrates with lower prices by focusing on conversion. Finally, bandits should be able to adjust price as long as targetable consumers are still available.

In lack of offline data that can be used for evaluation of targeting algorithms, such a system is powerful in picking up on contextual data should these be relevant (which we cannot know when we do not have relevant data from existing random trials to evaluate the algorithm – such data were only available from the A/B test for the app download decision point as shown in Section 5.4.3.2). The system presented here could, e.g., also be applied to pricing decisions for a new durable product that is launched globally where initial prices are set based on geolocation information and then price is adjusted per geolocation based on newly arriving sales data. Also, other data, e.g., on weather, news events and from social media such as Wikipedia or Twitter, could be taken into account. If managers are unwilling to give up agency on pricing decisions, their new "artificial colleague" can initially only provide recommendations. Once these have proven themselves to be reasonable, the learner can exert increasingly greater agency on price decisions.

Reports on applications of bandits and reinforcement learning more widely to business decisions are rare (Misra et al. 2019; Schwartz et al. 2017). Most studies focus on problems that lend themselves to automation and have not traditionally been addressed by human decision makers at scale (Li et al. 2010). Pricing is an area that tends to receive a lot of managerial attention. In this regard, this study shows how managerial intuition and substantive knowledge can be valuable assets in devising systems for data-driven optimization. Not only do they provide guidance,

e.g., for how to overcome cold start issues (Bietti et al. 2018; Padilla and Ascarza 2019; Loupos et al. 2019, also see Sections 5.4.4.1 and 5.4.3.3), but taking them into account is paramount in creating managerial acceptance for artificial intelligent systems.

5.5.4 Policy implications

From a wider societal perspective, this study speaks to issues of data-driven optimization and fairness: A main reason firms refrain from wider reaching price discrimination and pay explicit attention to fairness considerations, may be consumers' self-organization on forums and chat apps and the risk of customer protests and backlash should they feel treated unfairly (Odlyzko 2004; Chatterjee and McGinnis 2010; Li et al. 2019) – also see free-form comments on the survey among managers shown in Appendix A1, e.g., "the impact [of price personalization] was hugely negative within the community" (question 15 on page 159). It could seem desirable for policy makers not to rely on such self-organization, but to limit price discrimination and dynamic pricing practices by legal regulatory means. Especially when consumers face a monopolistic seller (Petro 2019), their self-organization and complaints may become ineffective and intervention by policy makers may be necessary to achieve a fair outcome.

It should be noted that such discriminatory practice does not solely benefit the firm, but can also be advantageous to consumers. Price discrimination will usually tend to charge higher prices to consumers with higher willingness-to-pay. To the extent that such higher willingness-to-pay derives from higher income and wealth, price discrimination can have a desirable redistributive effect. In the case of online games, high willingness-to-pay may, however, sometimes derive from addictive tendencies (Kwon et al. 2016; Nevskaya and Albuquerque 2019; Jo et al. 2020) rather than be reflective of personal wealth. Proactive regulation could hence be well advised in this setting.

A further consideration are policy measures to ensure healthy habits in the consumption of online games. Freemium pricing and mobile devices have given rise to an unparalleled increase in demand for online content by lowering entry and access barriers, bringing increased attention to issues of excessive use and online gaming addiction (European Commission 2014; Kwon et al. 2016; Jo et al. 2020). This study's findings suggest that pricing of premium upgrades and in-app purchases may not be an effective policy lever to impact spending sprees of heavy users in mobile games. Along these lines, measures such as excise taxes that intend to increase the price of addictive goods to curb their consumption may also be of mixed effectiveness (Wang et al. 2016). Overall revenue generation appears largely driven by design factors beyond pricing of in-app purchases or "gateway" offers (Schütze 2014). Usage limitations (Hahn et al. 2010; Nevskaya and Albuquerque 2019) and prohibition of certain design elements such a "Gatcha" and lottery-based purchases (Koeder and Tanaka 2017) may be more promising policy measures.

5.5.5 Limitations and future research

This study focuses on in-depth field experimentation in a popular game that is representative of the mobile game category (Levitt et al. 2016), leading to both high internal and external validity of the findings. External validity is further ensured by providing robust theoretical foundation for experiments and findings. Generally, a field experimentation approach entails high external validity, but evidence from further games would be helpful in supporting generalizability. Data available for analysis is representative of the situation of app developers. Platform operators such as Google and Apple have much more in-depth data that would likely make personalization of price and other treatments more effective. For good reason, especially to guard privacy and prevent discriminatory practice, this data is not available more widely though.

Also speaking to privacy issues, the authors cannot make user-level data available

for public use. Replication is still possible, but interested researchers will need to reach out to the corresponding author who can facilitate a data sharing and non-disclosure agreement with the data sponsor. Such measures are necessary due to the General Data Protection Regulation (GDPR) that applies to any company with customers in the area of the European Union. While this is a limitation of desirable and well intended research aims, such limitations on transparency are necessary if consumer data protections is to be ensured.

A further limitation is that the bandit-based learning system did not use contextual data in its decisions. While it was still helpful in devising a profitable skimming policy based on the randomization it introduced to the data while still "exploiting" based on a promising prior on 50% of users, applications of this system where contextual data matter and help to devise a further reaching personalization are a viable avenue for future research. Generally, further applications of reinforcement learning to marketing problems will be helpful in creating acceptance of such approaches in practice. An interesting extension of the current research in freemium settings could, e.g., personalize advertising load and content to different users based on their contextual data.

Finally, the analysis in this paper disregards competitive effects. While these may be low once a user downloaded an app and is "hooked" on its free version, studying these can make for an interesting extension in the setting of mobile apps and freemium software more generally, similar to recent work by Dubé et al. (2017a).

5.6 Conclusion

The present study is motivated from a substantive real-world problem that existing literature provides mixed guidance on: How should marketers set the price for "starter packs" in mobile games? In light of a long stream of literature cautioning against the use of low-price approaches (Lattin and Bucklin 1989; Blattberg

et al. 1995; Mela et al. 1997; Dekimpe et al. 1998; Jedidi et al. 1999; Anderson and Simester 2004), current managerial practice appears possibly flawed despite strong conceptual arguments in its favor; see 5.2.2.1 for an overview. Addressing this substantive issue, the study shows that managers in the $80 billion (2018 revenue) mobile gaming market can achieve higher profits by initially charging prices for in-app purchase promotions that are well beyond their comfort zone: (1) Only charge high prices to select user segments with high expected spending, (2) drop prices after a brief initial skimming period to entice users with lower willingness-to-pay to make a purchase, and (3) do so before users with lower willingness-to-pay disengage. These insights will similarly apply to pricing of other freemium digital content, e.g., subscriptions on a news website, in dating and networking apps and even in office and collaboration software. The study further showcases how field experimentation and a bandit-based online learning system can support the learning of profitable skimming price paths. The presented learning system is akin to an artificial pricing managers who aspires to implement a profitable skimming policy in concatenating several bandit learners with different reward specifications (see Figure 5.8 and Section 5.4.4.2). The authors believe that this learning system can inspire and inform profitable applications well beyond mobile game and freemium settings, e.g., in pricing news content, dating apps, other software-as-a-service solutions, or even during the global launch of a new durable product by learning profitable price paths in test markets.

5.7 Appendix

Appendix A1: Survey among mobile game managers

The survey was implemented in Google Forms and distributed on Mobiledevmemo (https://mobiledevmemo.com) – one of the largest mobile marketing blogs – and on Deconstructor of Fun (https://www.deconstructoroffun.com) – the leading mobile game design blog. Participation was incentivized with a lottery of a $30 Amazon voucher. The order of response options on all questions was randomized. 54 managers participated; 14 data scientists / analysts, 16 product managers / designers and 14 marketing managers; 10 respondents chose "Other" as a response. The full survey is available at `https://forms.gle/P2CPuLe9TH4bhh9E8` for readers' reference.

Table 5.9: Results of a survey among mobile game managers

Question 1: *How many years have you worked with mobile apps?* **Response:** 6.9 years (median: 2.4)
Question 2: *What types of apps have you worked on?* **Response:** 90.8% have worked in games; 66.7% have worked only in games. 9.3% have worked on other apps (ride hailing, news, lifestyle, social media, traveling) only.
Question 3: *Did the freemium apps that you worked on offer one-off purchases, subscriptions or both?* **Response:** 53.7% of respondents worked on apps that combined one-off purchases and subscription(s), 42.6% on apps that only offered one-off purchases, the rest on apps that only offered subscriptions.
Question 4: *How many years of professional marketing training do you have?* **Response:** 44.4% have zero years professional marketing training. The rest has an average of 3.2 years of such training (overall average 1.8 years).
Question 5: *Which role did you work in most?* **Response:** 16 respondents worked mostly in product design or management, 14 worked mostly in marketing, another 14 mostly in data science or analytics. 10 respondents did not provide an answer.
Question 6: *Monetizing and retaining app users can be challenging. In your opinion, to successfully monetize a freemium app's user base, you need to offer promotions and deals to users.* **Response:** 96.3% of respondents pick this over not using promotions and deals.

(continues on next page)

Continue Table 5.9: Results of a survey among mobile game managers

Question 7: *When new users download an app, do you think it is more important to...*

Response: 85.2%: focus on making them an attractive offer to get them to make a purchase quickly.
14.8%: focus on making them an offer at the highest possible price they are willing to pay.

Question 8: *Building on the previous question, which of the following statements is more correct in your opinion:*

Response: 61.1%: It is safer to sell a premium upgrade at a price lower than what a user is willing to pay – you can sell them more later.
38.9%: It is crucial to not set too low a price as a low price can impact what users are willing to pay in the future.

Question 9: *If you had the choice, which monetization configuration would you prefer in an app (assuming advertising revenue is the same):*

Response: 5.5%: Average lifetime value of users in the app is 10 USD; 1% of users spend 1000 USD each.
81.5%: Average lifetime value of users in the app is 10 USD; 10% of users spend 100 USD each.
13%: I'm indifferent between both options.

Question 10: *Which price (after discounts) do you think is appropriate for an initial promotion, i.e., an offer to new users, in the app?*

Response: <= 3 USD: 25.9%
3.01 to 5 USD: 33.3%
5.01 to 10 USD: 37.1%
> 10 USD: 3.7%

Question 11: *Retention and engagement are crucial to build an app's user base. In your opinion, when a user makes a purchase,...*

Response: 18.5%: it is because they are engaged with the app.
81.5%: it is because they are engaged, and it increases their engagement with the app.
0%: it increases their engagement with the app.

Question 12: *If you like, please share further thoughts how to best monetize a freemium app's user base:*

Response:

- by introducing time-limited events/sales/offers.
- really depends on region, how is your purchase screen set and also what features are paid and what are free.
- From a UA standpoint, I obviously prefer stronger early monetisation but it needs to be balanced so it doesn't cannibalise the overall LTV. One of the easiest monetisation tricks can be to try to segment the users on a day 0 based on their early signals and user properties and then adjust the starter packs and FTUE according to these segmentation.
- Through segmented promotions and personal offers based on segmentations. And a second currency with lives/ energy mechanics and rewarded video to decrease the waiting time.
- Hook them with great free content so that they want more and purchase.
- In general by not focusing too much on monetization strictly. The biggest mistake game teams make is focusing on the "supply" side: how you are selling the stuff you sell instead of the "demand" side: how to make players want more of the stuff you sell. There are some wins on the supply side sure, but big wins only come from big lifts in demand.
- create a perception of value for your target audience; communicate clearly the value (what's in it for them?) of what you have to sell; connect the items you sell with the player's goal in the game.

(continues on next page)

Continue Table 5.9: Results of a survey among mobile game managers

- The first purchase *has* to be seen by the user as "worth it." If you're pushing too much or things the user later realises are "not good enough," then you've scarred them for the rest of their app-lifetime.
- The cost of virtual goods is based on perception of value, and so before offering a user anything, it is imperative that you first establish the value of the thing you want to sell them. This can be in the form own utility or rarity or a combination of both. The power of the perceived value determines the money you can make from the item.
- Define the chase based on user's motivation (can be different by segment), the app will have a compelling value proposition and price point eventually. Run experiments to optimize price points.
- Spend more time thinking about what will make them enjoy the game, than how you can get money out of them. If they love the game, they will pay for things happily.
- I think it's important to have a healthy balance of free and premium users. Obviously we want all our users to become premium, but that's highly unlikely. Therefore, make sure that the free tier is also attractive for users. So keep investing in the free tier as well.
- These are hard questions. I don't feel confident about my answers to any of them.

Question 13: *Many apps sell premium upgrades at different prices in different countries. Price personalization is the practice of setting different prices for individual users based on further characteristics, e.g., the device they downloaded the app on. In your opinion, such price personalization is...*

Response: 70.3%: an essential tool in freemium apps to increase user monetization.
27.8%: more risky than useful in freemium apps.
1.9%: N/A

Question 14: *Have you used price personalization beyond country-based pricing (either through promotional offers or in-app purchase prices directly) in the apps you worked on?*

Response: 44.4%: No
53.7%: Yes
1.9%: N/A

Question 15: *If you like, please share further thoughts on price personalization:*

Response:

- offer personalization – yes; price personalization – yes, but the impact was hugely negative within the community.
- We use user behaviour to personalise the IAP prices during the special events.
- I think that it worth to an test price segments per country, even though on my experience it wasn't successful.
- At my previous job, we had a bigger user base on Android than on iOS. But they generated roughly the same revenue. Therefore we experimented with lower prices (which wasn't possible anyway in the App Store with their fixed price tier system) and it increased the number of buyers and later the total revenue as well.
- It can work well.
- Machine leaning can be used for classifying user types for price personalization as segmentation based tools (mainly used for adjusting prices in the aviation industry as for the case of discounters like Ryanair) might be unsuccessful in modeling the user's preference.
- I have never seen country based pricing generate a higher average revenue per user. I don't deem it risky, I just don't see the value in the extra work as you get the same return. However it is a way to gain featuring, which then justifies the additional effort.
- Do not discriminate players based on weird selection, offer good value to everyone.

(continues on next page)

Continue Table 5.9: Results of a survey among mobile game managers

– I think price personalization is essential. Some demographics have less to spend (think students, 65+, families, etc) and tailoring prices to them helps retaining users. Most of these concepts are already socially accepted, so people tend to accept them (think lower prices to museums, public transportation, etc)
– We set prices uniformly for all users, in all countries.
– If you value growing a healthy community around your app (which is very important for games), my bet is that you have more to lose than gain from price personalization, simply due to the fact that players talk to each other and it's (in my opinion rightly) perceived as unfair.

Appendix A2: Code for algorithm offline evaluation

This appendix presents an example of the code used for offline evaluation of Vowpal Wabbit's (contextual) bandit algorithm. The online learning system described in Section 5.7 uses the C++ implementation of Vowpal Wabbit. For this offline evaluation, the Python wrapper of Vowpal Wabbit was used – see `https://github.com/VowpalWabbit/vowpal_wabbit` for extensive documentation. The full code example also shows what other Python packages were used in this analysis and is available at `https://colab.research.google.com/drive/1ulPZHu2WkD2NwLTrwAl5Vp3917aGQwm-?usp=sharing`. The following code example shown here implements the policy estimation approach described in Section 5.3.2.2 and presented in Hitsch and Misra (2018) in more detail; the same code was used to generate the figures shown in Figure 5.6 on page 135:

```
# define number of arrivals to use for training
n = 100000

# set cost that contextual bandit is to minimize
df['cost'] = df['d1_offer_conversion']*(-1)
cost_kpi = 'cost'

# use first N arrivals to learn (data is sorted by user arrival time)
learn_df = df.iloc[0:n,:]

# use remainder for policy evaluation
deploy_df = df.iloc[n:len(df),:]

## add index to learn df
learn_df['index'] = range(0, len(learn_df))
learn_df = learn_df.set_index("index")

## add index to deploy df
deploy_df['index'] = range(0, len(deploy_df))
deploy_df = deploy_df.set_index("index")

# create model - this stores the model parameters in the python vowpal wabbit
    object
## model is set to consider three actions and be deployed with 0 learning rate
vw = pyvw.vw("--cb_explore 3 --epsilon 0")

# use the learn method to train the vw model (train model row by row)
i=0
for i in learn_df.index:
  ## provide data to cb in requested format
  action = learn_df.loc[i, "action"]
  cost = learn_df.loc[i, "cost"]
  probability = learn_df.loc[i, "probability"]
  feature1 = learn_df.loc[i, "country_groups"]
  feature2 = learn_df.loc[i, "device_groups"]
  ## do the actual learning
```

```
  vw.learn(str(action)+":"+str(cost)+":"+str(probability)+" |"+str(feature1)+" "+
      str(feature2))

# evaluate on remaining arrivals using 30 50% bootstraps
## for 2.99 offer = action 1, 4.99 offer = action 2, 29.99 offer = action 3
## with action 2 (best unpersonalized offer) as comparison baseline

# create objects to store results
reward_lift_best_t = np.zeros((30,1))
reward_lift_best_t[0] = 99
cost_lift_best_t = np.zeros((30,1))
cost_lift_best_t[0] = 99
assigned_action1 = np.zeros((30,1))
assigned_action1[0] = 99
assigned_action2 = np.zeros((30,1))
assigned_action2[0] = 99
assigned_action3 = np.zeros((30,1))
assigned_action3[0] = 99
same_assignment = np.zeros((30,1))
same_assignment[0] = 99
action_device_tier1 = np.zeros((30,1))
action_device_tier1[0] = 99
action_device_tier2 = np.zeros((30,1))
action_device_tier2[0] = 99
action_device_tier3 = np.zeros((30,1))
action_device_tier3[0] = 99
action_device_tier4 = np.zeros((30,1))
action_device_tier4[0] = 99
action_device_tier5 = np.zeros((30,1))
action_device_tier5[0] = 99

# start loop for 50 bootstrap iterations
k=1
for k in range(0,30):

  # use train test split data to bootstrap results
  from sklearn import datasets
  from sklearn.model_selection import train_test_split

  # get random 50% as test set
  drop_df, test_df = train_test_split(deploy_df, test_size=0.5, random_state=k)

  # add index to deploy df
  test_df['index'] = range(0, len(test_df))
  test_df = test_df.set_index("index")

  # predict row by row and output results
  prob1_vec = np.zeros((test_df.shape[0],1))
  prob1_vec[0] = 99
  prob2_vec = np.zeros((test_df.shape[0],1))
  prob2_vec[0] = 99
  prob3_vec = np.zeros((test_df.shape[0],1))
  prob3_vec[0] = 99

  j=0
  for j in test_df.index:
    feature1 = test_df.loc[j, "country_groups"]
    feature2 = test_df.loc[j, "device_groups"]
    vw_predict = vw.predict("| "+str(feature1)+" "+str(feature2))
    prob1_vec[j] = vw_predict[0]
    prob2_vec[j] = vw_predict[1]
    prob3_vec[j] = vw_predict[2]

  prob1_df = pd.DataFrame(prob1_vec)
  prob2_df = pd.DataFrame(prob2_vec)
  prob3_df = pd.DataFrame(prob3_vec)

  result_df = test_df.join(prob1_df)
  result_df.rename(columns={result_df.columns[57]: 'prob_action1'}, inplace=True)
  result_df = result_df.join(prob2_df)
```

```
    result_df.rename(columns={result_df.columns[58]: 'prob_action2'}, inplace=True)
    result_df = result_df.join(prob3_df)
    result_df.rename(columns={result_df.columns[59]: 'prob_action3'}, inplace=True)
    result_df

    # roll dice based on vowpal wabbit output probabilities to choose action
    def choose_action (row):
      c=np.random.choice(
        [1, 2, 3],
        1,
        p=[round(row['prob_action1'],10), round(row['prob_action2'],10), (1-round(row
            ['prob_action1'],10)-round(row['prob_action2'],10))])
      return c.astype(int)

    result_df['chosen_action'] = np.nan
    result_df['chosen_action'] = result_df.apply(lambda row: choose_action (row),axis
        =1)

    # identify overlaps between decision and actual assignment for offline evaluation
    def label_eval (row):
      if row['action'] == row['chosen_action'] :
        return 1
      return 0

    result_df['eval_label'] = result_df.apply(lambda row: label_eval (row),axis=1)

    # calculate expected average policy reward and cost based on overlapping
        instances
    reward_policy = result_df[result_df['eval_label']==1][reward_kpi].mean()
    cost_policy = result_df[result_df['eval_label']==1][cost_kpi].mean()

    reward_actions = result_df.groupby(['action'])[reward_kpi].mean()
    reward_best_t = reward_actions[2]
    reward_lift_best_t[k] = (reward_policy-reward_best_t)/reward_best_t
    cost_actions = result_df.groupby(['action'])[cost_kpi].mean()
    cost_best_t = cost_actions[2]
    cost_lift_best_t[k] = (cost_policy-cost_best_t)/cost_best_t

    # store decisions and number of overlapping decisions
    assigned_action1[k] = result_df[result_df['chosen_action']==1]['chosen_action'].
        count()
    assigned_action2[k] = result_df[result_df['chosen_action']==2]['chosen_action'].
        count()
    assigned_action3[k] = result_df[result_df['chosen_action']==3]['chosen_action'].
        count()
    same_assignment[k] = result_df['eval_label'].sum()

    action_device_tier1[k] = result_df[result_df['device_groups']=='Device tier 1']['
        chosen_action'].mean()
    action_device_tier2[k] = result_df[result_df['device_groups']=='Device tier 2']['
        chosen_action'].mean()
    action_device_tier3[k] = result_df[result_df['device_groups']=='Device tier 3']['
        chosen_action'].mean()
    action_device_tier4[k] = result_df[result_df['device_groups']=='Device tier 4']['
        chosen_action'].mean()
    action_device_tier5[k] = result_df[result_df['device_groups']=='Device tier 5']['
        chosen_action'].mean()

# aggregate results and prepare for visualization
reward_lift_best_t_df = pd.DataFrame(reward_lift_best_t)
reward_lift_best_t_df.rename(columns={reward_lift_best_t_df.columns[0]: 'Over Offer
    4.99'}, inplace=True)
reward_lift_df = reward_lift_best_t_df
cost_lift_best_t_df = pd.DataFrame(cost_lift_best_t)
cost_lift_best_t_df.rename(columns={cost_lift_best_t_df.columns[0]: 'Over Offer
    4.99'}, inplace=True)
cost_lift_df = cost_lift_best_t_df

assigned_action1_df = pd.DataFrame(assigned_action1)
assigned_action1_df.rename(columns={assigned_action1_df.columns[0]: 'Assigned to
```

```
    Offer␣2.99'}, inplace=True)
assigned_action2_df = pd.DataFrame(assigned_action2)
assigned_action2_df.rename(columns={assigned_action2_df.columns[0]: 'Assigned␣to␣
    Offer␣4.99'}, inplace=True)
assigned_action3_df = pd.DataFrame(assigned_action3)
assigned_action3_df.rename(columns={assigned_action3_df.columns[0]: 'Assigned␣to␣
    Offer␣29.99'}, inplace=True)
same_assignment_df = pd.DataFrame(same_assignment)
same_assignment_df.rename(columns={same_assignment_df.columns[0]: 'Overlapping␣
    assignment'}, inplace=True)

assignment_df = assigned_action1_df.join(assigned_action2_df)
assignment_df = assignment_df.join(assigned_action3_df)
assignment_df = assignment_df.join(same_assignment_df)

action_device_tier1_df = pd.DataFrame(action_device_tier1)
action_device_tier1_df.rename(columns={action_device_tier1_df.columns[0]: 'Device␣
    tier␣1'}, inplace=True)
action_device_tier2_df = pd.DataFrame(action_device_tier2)
action_device_tier2_df.rename(columns={action_device_tier2_df.columns[0]: 'Device␣
    tier␣2'}, inplace=True)
action_device_tier3_df = pd.DataFrame(action_device_tier3)
action_device_tier3_df.rename(columns={action_device_tier3_df.columns[0]: 'Device␣
    tier␣3'}, inplace=True)
action_device_tier4_df = pd.DataFrame(action_device_tier4)
action_device_tier4_df.rename(columns={action_device_tier4_df.columns[0]: 'Device␣
    tier␣4'}, inplace=True)
action_device_tier5_df = pd.DataFrame(action_device_tier5)
action_device_tier5_df.rename(columns={action_device_tier5_df.columns[0]: 'Device␣
    tier␣5'}, inplace=True)

device_action_df = action_device_tier1_df.join(action_device_tier2_df)
device_action_df = device_action_df.join(action_device_tier3_df)
device_action_df = device_action_df.join(action_device_tier4_df)
device_action_df = device_action_df.join(action_device_tier5_df)
```

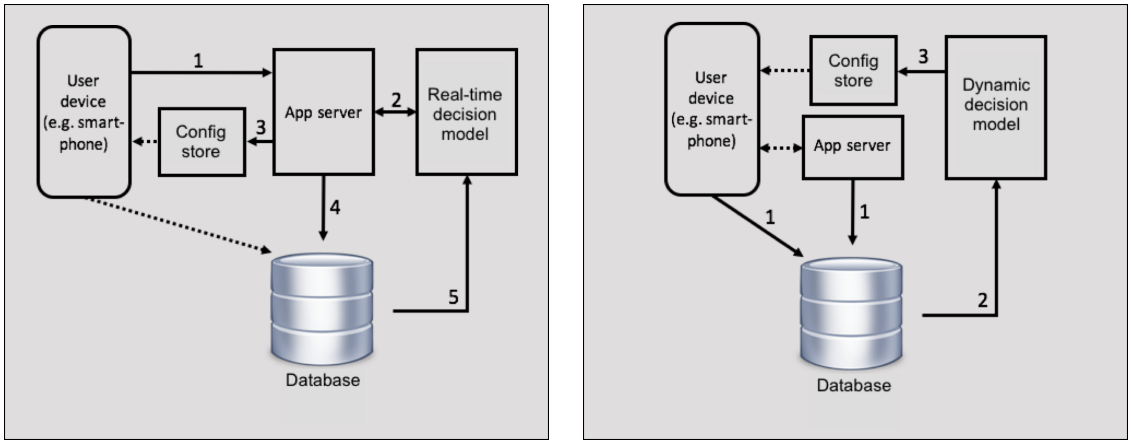
Appendix A3: Technical implementation of the online learning system

The online learning system used for field experimentation was built in collaboration with the data sponsor's engineering team. It consists of two sub-systems: (1) A real-time system that can make decisions at app download based on a provided model (either a trained contextual bandit or a heuristic); (2) a "downstream" learning system that has access to all behavioral data recorded in the company's logs and can use these for making decisions using flexible models.

The first sub-system (1) covers real-time personalization of the "starter pack" at app download. The left panel of Figure 5.11 shows a schematic depiction of the related decision architecture. A user device downloads the app from an app store. Upon the first launch of the app, the user device authenticates with the app server (step 1 in Figure 5.11) that in turn asks the real-time personalization service for a personalization decision, also sending relevant contextual data with the request (step 2). Based on the stored model's decision the app server then tells the content configuration store what configuration the user should receive (step 3). The configuration store continuously keeps the user device (=app client) appraised of this configuration assignment. The app server also sends data describing the personalization decision to a database (step 4) where the data are recorded in a well-behaved form. Both the app client and server continuously send data to this database to provide analysts and data scientists with users' behavioral traces. These are the base for model building by data scientists and for automated reinforcement of the real-time decision model (step 5).

The previous example focused on app download as a decision point and provides on-demand personalization in real-time. The right panel of Figure 5.11 shows how the "downstream" (i.e., after app download) user experience can then be adjusted by telling the config store to deliver a different configuration (e.g., with a different "starter pack") to a user device. App server and client keep sending behavioral traces

Figure 5.11: Schematic depiction of the architecture of the online learning system



Notes: Sub-system (1) on the left uses a stored model (either a trained contextual bandit or a heuristic) and device and country segments to make a decision in real-time when a new user downloads the app. Sub-system (2) on the right has access to detailed tracking data, including behavioral data, as stored in the company's logs and can use flexible models trained on this data to make decisions on the "downstream" user experience after app download.

to the database (step 1), generating a large pool of data. All this data can then be used to train dynamic decision models that can call into the config store to tell it to give a new configuration to a user device (step 3). As before, the config store remains in a constant loop with the client device to deliver the appropriate configuration. In the present study, this workflow is used to re-assign users to a different "starter pack" based on Vowpal Wabbit's bandit module or a different logic.

Appendix A4: Additional analysis of online learning run

To better understand why the bandits chose uniform actions across contexts, I use least absolute shrinkage and selection operator (LASSO – Santosa and Symes 1986; Tibshirani 1996) regression to analyze the individual price decisions on different days after app download (as shown in Figure 5.8) to see what explanatory variables are retained and if meaningful patterns emerge.

I use LASSO regression as implemented in the R package glmnet for variable selection and regularization. The regularization parameter Lambda is tuned using grid search and five-fold cross-validation. Table 5.10 shows results of this analysis. Models are underpowered in that the price decision variable is dropped from four of them. For the two models where the price decision is retained (model M8.1 and model M8.2), the recommended decision is largely in line with regression results in Section 5.4.4.2. Coefficients on contextual variable main and interaction effects are inconsistently retained and overall very small, explaining why the bandits ignored contextual variables in their decisions. Note that the sample size decreases for later decision points (in days after app download) as the bandit only makes a decision for users that are active in the two days before the decision point (see Figure 5.8) and have not made a purchase yet. While the result that contextual variables do not matter to the reward-action distribution is disappointing, this online run of the bandit system is successful in indicating a viable uniform skimming policy as laid out in Section 5.4.4.2.

Table 5.10: Regression analysis of reward on contextual variables

Model	M8.1	M8.2	M8.3	M8.4	M8.5	M8.6
Decision point	Day 2	Day 4	Day 6	Day 8	Day 10	Day 12
Reward (= y)	Offer revenue	Offer revenue	Offer revenue	Offer conversion	Offer conversion	Offer conversion
Price decision						
Keep price (1)	−0.006	0.0282	Dropped	Dropped	Dropped	Dropped
Device tier (Tier 5 excluded as reference category)						
Tier 1	0.028	−0.0232	0.0056	Dropped	Dropped	Dropped
Tier 2	Dropped	−0.0522	Dropped	Dropped	Dropped	Dropped
Tier 3	−0.0003	−0.0532	Dropped	Dropped	Dropped	Dropped
Tier 4	−0.0007	0.0055	Dropped	Dropped	Dropped	Dropped
Country tier (Tier 4 excluded as reference category)						
Tier 1	0.0169	−0.0091	Dropped	Dropped	Dropped	Dropped
Tier 2	0.0136	Dropped	Dropped	Dropped	Dropped	Dropped
Tier 3	−0.0063	−0.0076	−0.0005	Dropped	Dropped	Dropped
Contextual variables						
Sessions	0.0057	−0.0000	0.0041	Dropped	0.0004	0.0041
Rounds	−0.0005	Dropped	0.0010	0.0006	Dropped	Dropped
In-app transactions	−0.0003	0.0017	0.0106	0.0005	Dropped	Dropped
Gift claims	0.0033	0.0001	0.0008	Dropped	0.0002	Dropped
Gifts sent	0.0002	Dropped	Dropped	Dropped	Dropped	−0.0001
Ad views	0.002	−0.0006	0.0004	Dropped	0.0008	Dropped
Thanks to friends	0.0031	−0.0001	Dropped	Dropped	0.0005	Dropped
Messages	−0.0026	Dropped	Dropped	Dropped	−0.0003	0.0004
Enemies beaten	Dropped	−0.0000	Dropped	Dropped	Dropped	0.0003
Interactions between price decision and contextual variables						
Device tier 1	−0.0223	−0.0379	Dropped	Dropped	Dropped	Dropped
Device tier 2	Dropped	−0.003	Dropped	Dropped	Dropped	Dropped
Device tier 3	Dropped	−0.0196	Dropped	Dropped	Dropped	Dropped
Device tier 4	Dropped	Dropped	Dropped	Dropped	Dropped	Dropped
Country tier 1	−0.0017	Dropped	Dropped	Dropped	Dropped	Dropped
Country tier 2	Dropped	0.0024	Dropped	Dropped	Dropped	Dropped
Country tier 3	−0.0001	−0.0148	Dropped	Dropped	Dropped	Dropped
Sessions	0.0072	−0.0061	−0.0058	Dropped	Dropped	−0.0037
Rounds	0.0020	0.0027	Dropped	Dropped	Dropped	Dropped
In-app transactions	−0.0011	Dropped	Dropped	Dropped	−0.0016	Dropped
Gift claims	−0.007	−0.0001	Dropped	Dropped	−0.0000	Dropped
Gifts sent	−0.0004	Dropped	Dropped	Dropped	−0.0000	Dropped
Ad views	−0.0031	−0.0000	Dropped	Dropped	Dropped	Dropped
Thanks to friends	0.0074	Dropped	Dropped	Dropped	−0.0006	Dropped
Messages	−0.0028	Dropped	Dropped	Dropped	Dropped	Dropped
Enemies beaten	−0.0008	Dropped	Dropped	Dropped	Dropped	0.0001
Intercept	−0.0183	0.0611	0.0056	0.0021	0.0009	0.0008
Lambda	0.001	0.001	0.005	0.004	0.001	0.001
N	47,033	29,755	19,488	21,259	14,256	8,388

Notes: Results for LASSO regressions of reward on decision indicator and contextual variables at different decision points.

Bibliography

Aboussalah, A. M. and C.-G. Lee (2020): "Continuous Control with Stacked Deep Dynamic Recurrent Reinforcement Learning for Portfolio Optimization," *Expert Systems with Applications*, 140, 112891.

Acquisti, A. and H. R. Varian (2005): "Conditioning Prices on Purchase History," *Marketing Science*, 24, 367–381.

Alaa, A. M., M. Weisz, and M. van der Schaar (2017): "Deep Counterfactual Networks with Propensity-Dropout," *arXiv preprint arXiv:1706.05966*, `https://arxiv.org/abs/1706.05966` (Accessed May 18 2020).

Alsén, A., J. Runge, A. Drachen, and D. Klapper (2016): "Play with Me? Understanding and Measuring the Social Aspect of Casual Gaming," in *Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Alter, A. (2017): *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*, London, United Kingdom: Penguin Books.

Anderson, C. (2009): *Free: The Future of a Radical Price*, New York City, New York: Random House.

Anderson, E. T. and D. I. Simester (2004): "Long-Run Effects of Promotion Depth on New Versus Established Customers: Three Field Studies," *Marketing Science*, 23, 4–20.

——— (2010): "Price Stickiness and Customer Antagonism," *The Quarterly Journal of Economics*, 125, 729–765.

Angermueller, C., T. Pärnamaa, L. Parts, and O. Stegle (2016): "Deep Learning for Computational Biology," *Molecular Systems Biology*, 12, 878.

Ansari, A. and C. F. Mela (2003): "E-Customization," *Journal of Marketing Research*, 40, 131–145.

App Annie (2018): "The State of Mobile 2019," *App Annie Blog*, `https://www.appannie.com/en/go/state-of-mobile-2019/` (Accessed May 18 2020).

Appel, G., B. Libai, E. Muller, and R. Shachar (2019): "On the Monetization of Mobile Apps," *International Journal of Research in Marketing*, 37, 93–107.

Arora, N., X. Dreze, A. Ghose, J. D. Hess, R. Iyengar, B. Jing, Y. Joshi, V. Kumar, N. Lurie, S. Neslin, et al. (2008): "Putting One-to-One Marketing to Work: Personalization, Customization, and Choice," *Marketing Letters*, 19, 305.

Arora, S., F. Ter Hofstede, and V. Mahajan (2017): "The Implications of Offering Free Versions for the Performance of Paid Mobile Apps," *Journal of Marketing*, 81, 62–78.

Ascarza, E. (2018): "Retention Futility: Targeting High-Risk Customers Might Be Ineffective," *Journal of Marketing Research*, 55, 80–98.

Ascarza, E., R. Iyengar, and M. Schleicher (2016): "The Perils of Proactive Churn Prevention Using Plan Recommendations: Evidence from a Field Experiment," *Journal of Marketing Research*, 53, 46–60.

Ascarza, E., A. Lambrecht, and N. Vilcassim (2012): "When Talk Is 'Free': The Effect of Tariff Structure on Usage Under Two- and Three-Part Tariffs," *Journal of Marketing Research*, 49, 882–899.

Ascarza, E., S. A. Neslin, O. Netzer, Z. Anderson, P. S. Fader, S. Gupta, B. G. Hardie, A. Lemmens, B. Libai, D. Neal, et al. (2018): "In Pursuit of Enhanced Customer Retention management: Review, Key Issues, and Future Directions," *Customer Needs and Solutions*, 5, 65–81.

Baker, W., D. Kiewell, and G. Winkler (2014): "Using Big Data to Make Better Pricing Decisions," *McKinsey Analysis*, `https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/using-big-data-to-make-better-pricing-decisions` (Accessed May 18 2020).

Balasubraman, S., R. A. Peterson, and S. L. Jarvenpaa (2002): "Exploring the Implications of M-Commerce for Markets and Marketing," *Journal of the Academy of Marketing Science*, 30, 348–361.

Banerjee, T., G. Mukherjee, S. Dutta, and P. Ghosh (2019): "A Large-Scale Constrained Joint Modeling Approach for Predicting User Activity, Engagement, and Churn With Application to Freemium Mobile Games," *Journal of the American Statistical Association*, 1–29.

Bapna, R., J. Ramaprasad, and A. Umyarov (2017): "Monetizing Freemium Communities: Does Paying for Premium Increase Social Engagement?" *Management Information Systems Quarterly (MISQ)*, 42, 719–735.

Bapna, R. and A. Umyarov (2015): "Do Your Online Friends Make You Pay? A Randomized Field Experiment on Peer Influence in Online Social Networks," *Management Science*, 61, 1902–1920.

Barton, D. and D. Court (2012): "Making Advanced Analytics Work for You," *Harvard Business Review*, 90, 78–83.

Baumann, A., S. Lessmann, K. Coussement, and K. W. De Bock (2015): "Maximize What Matters: Predicting Customer Churn With Decision-Centric Ensemble Selection," in *European Conference on Information Systems*.

Bawa, K. and R. Shoemaker (2004): "The Effects of Free Sample Promotions on Incremental Brand Sales," *Marketing Science*, 23, 345–363.

Becker, G. S., M. Grossman, and K. M. Murphy (1991): "Rational Addiction and the Effect of Price on Consumption," *American Economic Review*, 81, 237–241.

Becker, G. S. and K. M. Murphy (1988): "A Theory of Rational Addiction," *Journal of Political Economy*, 96, 675–700.

Bell, D. R. and J. Lattin (1998): "Shopping Behavior and Consumer Preference for Retail Price Format: Why "Large Basket" Shoppers Prefer EDLP," *Marketing Science*, 17, 66–88.

Bemmaor, A. C. and D. Mouchoux (1991): "Measuring the Short-Term Effect of In-Store Promotion and Retail Advertising on Brand Sales: A Factorial Experiment," *Journal of Marketing Research*, 28, 202–214.

Berger, P. D. and N. I. Nasr (1998): "Customer Lifetime Value: Marketing Models and Applications," *Journal of Interactive Marketing*, 12, 17–30.

Bertsimas, D. and A. J. Mersereau (2007): "A Learning Approach for Interactive Marketing to a Customer Segment," *Operations Research*, 55, 1120–1135.

Bettman, J. R. (1979): *Information Processing Theory of Consumer Choice*, Boston, Massachusetts: Addison-Wesley Publishing Company.

Bietti, A., A. Agarwal, and J. Langford (2018): "A Contextual Bandit Bake-Off," *arXiv preprint arXiv:1802.04064*, `https://arxiv.org/abs/1802.04064` (Accessed May 18 2020).

Bishop, C. M. (1995): "Training with Noise is Equivalent to Tikhonov Regularization," *Neural Computation*, 7, 108–116.

Blattberg, R. C., R. Briesch, and E. J. Fox (1995): "How Promotions Work," *Marketing Science*, 14, G122–G132.

Blattberg, R. C. and J. Deighton (1996): "Manage Marketing by the Customer Equity Test," *Harvard Business Review*, 74, 136.

Block, J. J. (2008): "Issues for DSM-V: Internet Addiction," *American Journal of Psychiatry*, 165, 306–307.

Bose, R. (2009): "Advanced Analytics: Opportunities and Challenges," *Industrial Management & Data Systems*, 109, 155–172.

Briesch, R. and P. Rajagopal (2010): "Neural Network Applications in Consumer Behavior," *Journal of Consumer Psychology*, 20, 381–389.

Bucklin, R. E., J. M. Lattin, A. Ansari, S. Gupta, D. Bell, E. Coupey, J. D. Little, C. Mela, A. Montgomery, and J. Steckel (2002): "Choice and the Internet: From Clickstream to Research Stream," *Marketing Letters*, 13, 245–258.

Bucklin, R. E. and C. Sismeiro (2009): "Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing," *Journal of Interactive Marketing*, 23, 35–48.

Burez, J. and D. Van den Poel (2007): "CRM at a Pay-TV Company: Using Analytical Models to Reduce Customer Attrition by Targeted Marketing for Subscription Services," *Expert Systems with Applications*, 32, 277–288.

——— (2009): "Handling Class Imbalance in Customer Churn Prediction," *Expert Systems with Applications*, 36, 4626–4636.

Calder, B. J., E. C. Malthouse, and U. Schaedel (2009): "An Experimental Study of the Relationship between Online Engagement and Advertising Effectiveness," *Journal of Interactive Marketing*, 23, 321–331.

Carter, B. (2019): "Freemium Conversion Issues? Why You Need to Address the Penny Gap," *GoSquared Blog*, `https://www.gosquared.com/blog/freemium-conversion-issues` (Accessed May 18 2020).

Chamberlain, B. P., A. Cardoso, C. H. Liu, R. Pagliari, and M. P. Deisenroth (2017): "Customer Lifetime Value Prediction Using Embeddings," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 1753–1762.

Chandon, P., J. W. Hutchinson, E. T. Bradlow, and S. H. Young (2009): "Does In-Store Marketing Work? Effects of the Number and Position of Shelf Facings on Brand Attention and Evaluation at the Point of Purchase," *Journal of Marketing*, 73, 1–17.

Chatterjee, P. and J. McGinnis (2010): "Customized Online Promotions: Moderating Effect of Promotion Type on Deal Value, Perceived Fairness, and Purchase Intent," *Journal of Applied Business Research (JABR)*, 26.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002): "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, 16, 321–357.

Chen, C. and L. Leung (2016): "Are You Addicted to Candy Crush Saga? An Exploratory Study Linking Psychological Factors to Mobile Social Game Addiction," *Telematics and Informatics*, 33, 1155–1166.

Chen, T., B. Sun, and V. Singh (2009): "An Empirical Investigation of the Dynamic Effect of Marlboro's Permanent Pricing Shift," *Marketing Science*, 28, 740–758.

Chen, Z.-Y., Z.-P. Fan, and M. Sun (2012): "A Hierarchical Multiple Kernel Support Vector Machine for Customer Churn Prediction Using Longitudinal Behavioral Data," *European Journal of Operational Research*, 223, 461–472.

Cole, H. and M. D. Griffiths (2007): "Social Interactions in Massively Multiplayer Online Role-Playing Gamers," *Cyberpsychology & Behavior*, 10, 575–583.

Coussement, K., S. Lessmann, and G. Verstraeten (2017): "A Comparative Analysis of Data Preparation Algorithms for Customer Churn Prediction: A Case Study in the Telecommunication Industry," *Decision Support Systems*, 95, 27–36.

Coussement, K. and D. Van den Poel (2008): "Churn Prediction in Subscription Services: An Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques," *Expert Systems with Applications*, 34, 313–327.

Crone, S. F., S. Lessmann, and R. Stahlbock (2006): "The Impact of Preprocessing on Data Mining: An Evaluation of Classifier Sensitivity in Direct Marketing," *European Journal of Operational Research*, 173, 781–800.

Dawes, R. M. and B. Corrigan (1974): "Linear Models in Decision Making," *Psychological Bulletin*, 81, 95–111.

De Haan, E., P. Kannan, P. C. Verhoef, and T. Wiesel (2018): "Device Switching in Online Purchasing: Examining the Strategic Contingencies," *Journal of Marketing*, 82, 1–19.

Deci, E. L., R. Koestner, and R. M. Ryan (1999): "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation," *Psychological Bulletin*, 125, 627–668.

Dekimpe, M. G., D. M. Hanssens, and J. M. Silva-Risso (1998): "Long-Run Effects of Price Promotions in Scanner Markets," *Journal of Econometrics*, 89, 269–291.

Dew, R. and A. Ansari (2018): "Bayesian Nonparametric Customer Base Analysis with Model-based Visualizations," *Marketing Science*, 37, 216–235.

Donkers, B., P. C. Verhoef, and M. G. de Jong (2007): "Modeling CLV: A Test of Competing Models in the Insurance Industry," *Quantitative Marketing and Economics*, 5, 163–190.

Du, R. Y. and W. A. Kamakura (2008): "Where Did All that Money Go? Understanding How Consumers Allocate Their Consumption Budget," *Journal of Marketing*, 72, 109–131.

Du, S., J. Lee, and F. Ghaffarizadeh (2019): "Improve User Retention with Causal Learning," in *The 2019 ACM SIGKDD Workshop on Causal Discovery*, 34–49.

Dubé, J.-P., Z. Fang, N. Fong, and X. Luo (2017a): "Competitive Price Targeting with Smartphone Coupons," *Marketing Science*, 36, 944–975.

Dubé, J.-P., X. Luo, and Z. Fang (2017b): "Self-Signaling and Prosocial Behavior: A Cause Marketing Experiment," *Marketing Science*, 36, 161–186.

Dubé, J.-P. and S. Misra (2017): "Scalable Price Targeting," *National Bureau of Economic Research Working Paper*, https://www.nber.org/papers/w23775.pdf (Accessed May 18 2020).

Dudík, M., J. Langford, and L. Li (2011): "Doubly Robust Policy Evaluation and Learning," *arXiv preprint arXiv:1103.4601*, https://arxiv.org/abs/1103.4601 (Accessed May 18 2020).

Dziurzynski, L., E. Wadsworth, P. Fader, E. M. Feit, D. McCarthy, B. Hardie, A. Gopalakrishnan, E. Schwartz, and Y. Zhang (2012): "Package BTYD," *CRAN*, https://cran.r-project.org/web/packages/BTYD/index.html (Accessed May 18 2020).

Einav, L., J. Levin, I. Popov, and N. Sundaresan (2014): "Growth, Adoption, and Use of Mobile E-Commerce," *American Economic Review*, 104, 489–94.

Einhorn, H. J. (1970): "The Use of Non-Linear, Non-Compensatory Models in Decision Making," *Psychological Bulletin*, 73, 221–230.

Eisingerich, A. B., A. Marchand, M. P. Fritze, and L. Dong (2019): "Hook vs. Hope: How to Enhance Customer Engagement through Gamification," *International Journal of Research in Marketing*, 36, 200–215.

Ekinci, Y., F. Ülengin, N. Uray, and B. Ülengin (2014): "Analysis of Customer Lifetime Value and Marketing Expenditure Decisions through a Markovian-Based Model," *European Journal of Operational Research*, 237, 278–288.

Elberg, A., P. Gardete, R. Macera, and C. Noton (2019): "Dynamic Effects of Price Promotions: Field Evidence, Consumer Search, and Supply-Side Implications," *Quantitative Marketing and Economics*, 17, 1–58.

Ellickson, P. B. and S. Misra (2008): "Supermarket Pricing Strategies," *Marketing Science*, 27, 811–828.

Ellickson, P. B., S. Misra, and H. S. Nair (2012): "Repositioning Dynamics and Pricing Strategy," *Journal of Marketing Research*, 49, 750–772.

Erdem, T., S. Imai, and M. P. Keane (2003): "Brand and Quantity Choice Dynamics Under Price Uncertainty," *Quantitative Marketing and Economics*, 1, 5–64.

Erdem, T., M. P. Keane, and B. Sun (2008): "A Dynamic Model of Brand Choice When Price and Advertising Signal Product Quality," *Marketing Science*, 27, 1111–1125.

European Commission (2014): "Commission and Member States to Raise Consumer Concerns with App Industry," *European Commission Press Releases*, https://ec.europa.eu/commission/presscorner/detail/en/IP_14_187 (Accessed May 18 2020).

Eyal, N. (2014): *Hooked: How to Build Habit-Forming Products*, London, United Kingdom: Penguin Books.

Fader, P. S., B. G. Hardie, and K. L. Lee (2005): "RFM and CLV: Using Iso-Value Curves for Customer Base Analysis," *Journal of Marketing Research*, 42, 415–430.

Foubert, B. and E. Gijsbrechts (2016): "Try It, You'll Like It–or Will You? The Perils of Early Free-Trial Promotions for High-Tech Service Adoption," *Marketing Science*, 35, 810–826.

Gameanalytics (2019): "Mobile Gaming Benchmarks," *Gameanalytics Blog*, `https://pages.gameanalytics.com/rs/686-EPV-320/images/H1-2019-Mobile-Benchmarks-Report-GameAnalytics.pdf` (Accessed May 18 2020).

Gerstner, E. (1985): "Do Higher Prices Signal Higher Quality?" *Journal of Marketing Research*, 22, 209–215.

Ghose, A., A. Goldfarb, and S. P. Han (2013): "How Is the Mobile Internet Different? Search Costs and Local Activities," *Information Systems Research*, 24, 613–631.

Ghose, A. and S. P. Han (2014): "Estimating Demand for Mobile Applications in the New Economy," *Management Science*, 60, 1470–1488.

Glady, N., B. Baesens, and C. Croux (2009): "Modeling Churn Using Customer Lifetime Value," *European Journal of Operational Research*, 197, 402–411.

Glorot, X., A. Bordes, and Y. Bengio (2011): "Deep Sparse Rectifier Neural Networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 315–323.

Gordon, B. R. and B. Sun (2015): "A Dynamic Model of Rational Addiction: Evaluating Cigarette Taxes," *Marketing Science*, 34, 452–470.

Gordon, B. R., F. Zettelmeyer, N. Bhargava, and D. Chapsky (2019): "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook," *Marketing Science*, 38, 193–225.

Gough, C. (2018): "Distribution of Digital Games Market Revenue Worldwide in 2017, by Monetization Model," *Statista - The Statistics Portal*, `https://www.statista.com/statistics/821451/distribution-digital-games-market-revenue-monetization-model/` (Accessed May 18 2020).

Gu, X., P. Kannan, and L. Ma (2018): "Selling the Premium in Freemium," *Journal of Marketing*, 82, 10–27.

Guhl, D., H. W. von Mohrenfels, J. Abshagen, and D. Klapper (2016): "Measuring Marketing Success: Estimating the Effect of Social Media and TV Advertising on Brand Attention," *Marketing: ZFP–Journal of Research and Management*, 38, 44–54.

Günter, T. and D. Klapper (2007): "Do the Long-Run Category Demand Effects of Retailer Promotions Vary across Different Store Types?" *Marketing-Journal of Research and Management*, 3, 17–33.

Gupta, S., D. R. Lehmann, and J. A. Stuart (2004): "Valuing Customers," *Journal of Marketing Research*, 41, 7–18.

Hadiji, F., R. Sifa, A. Drachen, C. Thurau, K. Kersting, and C. Bauckhage (2014): "Predicting Player Churn in the Wild," in *2014 IEEE Conference on Computational Intelligence and Games*, IEEE, 1–8.

Hahn, R. A., J. L. Kuzara, R. Elder, R. Brewer, S. Chattopadhyay, J. Fielding, T. S. Naimi, T. Toomey, J. C. Middleton, B. Lawrence, et al. (2010): "Effectiveness of Policies Restricting Hours of Alcohol Sales in Preventing Excessive Alcohol Consumption and Related Harms," *American Journal of Preventive Medicine*, 39, 590–604.

Halbheer, D., F. Stahl, O. Koenigsberg, and D. R. Lehmann (2014): "Choosing a Digital Content Strategy: How Much Should Be Free?" *International Journal of Research in Marketing*, 31, 192–206.

Hamari, J., N. Hanner, and J. Koivisto (2020): "Why Pay Premium in Freemium Services? A Study on Perceived Value, Continued Use and Purchase Intentions in Free-to-Play Games," *International Journal of Information Management*, 51, 10–24.

Hamari, J. and L. Keronen (2017): "Why Do People Buy Virtual Goods: A Meta-Analysis," *Computers in Human Behavior*, 71, 59–69.

Han, S. P., S. Park, and W. Oh (2015): "Mobile App Analytics: A Multiple Discrete-Continuous Choice Framework," *Management Information Systems Quarterly (MISQ)*, 40, 983–1008.

Heilman, C. M., F. Kaefer, and S. D. Ramenofsky (2003): "Determining the Appropriate Amount of Data for Classifying Consumers for Direct Marketing Purposes," *Journal of Interactive Marketing*, 17, 5–28.

Hendel, I. and A. Nevo (2003): "The Post-Promotion Dip Puzzle: What Do the Data Have to Say?" *Quantitative Marketing and Economics*, 1, 409–424.

Heyman, J. and D. Ariely (2004): "Effort for Payment: A Tale of Two Markets," *Psychological Science*, 15, 787–793.

Hitsch, G. J. and S. Misra (2018): "Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation," *SSRN Working Paper 3111957*, `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3111957` (Accessed May 18 2020).

Ho, T.-H., C. S. Tang, and D. R. Bell (1998): "Rational Shopping Behavior and the Option Value of Variable Pricing," *Management Science*, 44, 145–160.

Hoch, S., X. Dreze, and M. Purk (1994): "EDLP, Hi-Lo, and Margin Arithmetic," *Journal of Marketing*, 58, 16–27.

Hong, Y. and P. A. Pavlou (2014): "Product Fit Uncertainty in Online Markets: Nature, Effects, and Antecedents," *Information Systems Research*, 25, 328–344.

Hornik, K. (1991): "Approximation Capabilities of Multilayer Feedforward Networks," *Neural Networks*, 4, 251–257.

Horvitz, D. G. and D. J. Thompson (1952): "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Hu, M. Y. and C. Tsoukalas (2003): "Explaining Consumer Choice through Neural Networks: The Stacked Generalization Approach," *European Journal of Operational Research*, 146, 650–660.

Huang, J.-H., C.-T. Chang, and C. Y.-H. Chen (2005): "Perceived Fairness of Pricing on the Internet," *Journal of Economic Psychology*, 26, 343–361.

Jedidi, K., C. F. Mela, and S. Gupta (1999): "Managing Advertising and Promotion for Long-Run Profitability," *Marketing Science*, 18, 1–22.

Jo, W., S. Sunder, J. Choi, and M. Trivedi (2020): "Protecting Consumers from Themselves: Assessing Consequences of Usage Restriction Laws on Online Game Usage and Spending," *Marketing Science*, 39, 117–133.

Kaplan, O. (2019): "Mobile Gaming Is a 68.5 billion USD Global Business, and Investors Are Buying In," *Techcrunch*, `https://techcrunch.com/2019/08/22/mobile-gaming-mints-money/` (Accessed May 18 2020).

Katz, S. and A. Lavack (2002): "Tobacco Related Bar Promotions: Insights from Tobacco Industry Documents," *Tobacco Control*, 11, i92–i101.

Kim, Y., W. N. Street, G. J. Russell, and F. Menczer (2005): "Customer Targeting: A Neural Network Approach Guided by Genetic Algorithms," *Management Science*, 51, 264–276.

Kingma, D. P. and J. Ba (2014): "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, `https://arxiv.org/abs/1412.6980` (Accessed May 18 2020).

Koeder, M. J. and E. Tanaka (2017): "Game of Chance Elements in Free-to-Play Mobile Games: A Freemium Business Model Monetization Tool in Need of Self-Regulation?" in *28th European Regional Conference of the International Telecommunications Society (ITS) "Competition and Regulation in the Information Age"*.

Kosinski, M., D. Stillwell, and T. Graepel (2013): "Private Traits and Attributes are Predictable from Digital Records of Human Behavior," *Proceedings of the National Academy of Sciences*, 110, 5802–5805.

Kumar, A., V. R. Rao, and H. Soni (1995): "An Empirical Comparison of Neural Network and Logistic Regression Models," *Marketing Letters*, 6, 251–263.

Kumar, V. (2014): "Making 'Freemium' Work," *HBR.org*, https://hbr.org/2014/05/making-freemium-work (Accessed May 18 2020).

Kwon, H. E., H. So, S. P. Han, and W. Oh (2016): "Excessive Dependence on Mobile Social Apps: A Rational Addiction Perspective," *Information Systems Research*, 27, 919–939.

Lambrecht, A., A. Goldfarb, A. Bonatti, A. Ghose, D. G. Goldstein, R. Lewis, A. Rao, N. Sahni, and S. Yao (2014): "How Do Firms Make Money Selling Digital Goods Online?" *Marketing Letters*, 25, 331–341.

Lambrecht, A. and K. Misra (2016): "Fee or free: When Should Firms Charge for Online Content?" *Management Science*, 63, 1150–1165.

Lattin, J. M. and R. E. Bucklin (1989): "Reference Effects of Price and Promotion on Brand Choice Behavior," *Journal of Marketing Research*, 26, 299–310.

LeCun, Y., Y. Bengio, and G. Hinton (2015): "Deep Learning," *Nature*, 521, 436–444.

LeCun, Y., L. Bottou, G. B. Orr, and K.-R. Müller (1998): "Efficient Backprop," in *Neural networks: Tricks of the trade*, Berlin, Germany: Springer, 9–50.

Lee, C., V. Kumar, and S. Gupta (2017): "Designing Freemium: Strategic Balancing of Growth and Monetization," *SSRN Working Paper 2767135*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2767135 (Accessed May 18 2020).

Lehdonvirta, V. (2009): "Virtual Item Sales as a Revenue Model: Identifying Attributes that Drive Purchase Decisions," *Electronic Commerce Research*, 9, 97–113.

Lemmens, A. and C. Croux (2006): "Bagging and Boosting Classification Trees to Predict Churn," *Journal of Marketing Research*, 43, 276–286.

Lemmens, A. and S. Gupta (2020): "Managing Churn to Maximize Profits," *Marketing Science*, Forthcoming.

Lessmann, S. (2004): "Solving Imbalanced Classification Problems with Support Vector Machines." in *IC-AI*, vol. 4, 214–220.

Lessmann, S., B. Baesens, H.-V. Seow, and L. C. Thomas (2015): "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research," *European Journal of Operational Research*, 247, 124–136.

Lessmann, S. and S. Voss (2009): "A Reference Model for Customer-Centric Data Mining with Support Vector Machines," *European Journal of Operational Research*, 199, 520–530.

Levitt, S. D., J. A. List, S. Neckermann, and D. Nelson (2016): "Quantity Discounts on a Virtual Good: The Results of a Massive Pricing Experiment at King Digital Entertainment," *Proceedings of the National Academy of Sciences*, 113, 7323–7328.

Lewis, C., N. Wardrip-Fruin, and J. Whitehead (2012): "Motivational Game Design Patterns of 'ville Games," in *Proceedings of the International Conference on the Foundations of Digital Games*, 172–179.

Li, H. A., S. Jain, and P. Kannan (2018): "Optimal Design of Content Samples for Digital Products and Services," *Journal of Marketing Research*, 56, 419–438.

Li, L., W. Chu, J. Langford, and R. E. Schapire (2010): "A Contextual-Bandit Approach to Personalized News Article Recommendation," in *Proceedings of the 19th International Conference on World Wide Web*, ACM, 661–670.

Li, X., K. J. Li, and X. Wang (2019): "Transparency of Behavior-Based Pricing," *Journal of Marketing Research*, 57, 78–99.

Loupos, P., A. Nathan, and M. Cerf (2019): "Starting Cold: The Power of Social Networks in Predicting Non-Contractual Customer Behavior," *SSRN Working Paper 3001978*, `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3001978` (Accessed May 18 2020).

Malthouse, E. C. and R. C. Blattberg (2005): "Can We Predict Customer Lifetime Value?" *Journal of Interactive Marketing*, 19, 2–16.

Martinez, M. (2014): "Amazon Error May End 'Dynamic Pricing'," *ABC News*, `https://abcnews.go.com/Technology/story?id=119399&page=1&page=1` (Accessed May 18 2020).

Matz, S., M. Kosinski, G. Nave, and D. Stillwell (2017): "Psychological Targeting as an Effective Approach to Digital Mass Persuasion," *Proceedings of the National Academy of Sciences*, 114, 12714–12719.

McCulloch, W. S. and W. Pitts (1943): "A Logical Calculus of the Ideas Immanent in Nervous Activity," *The Bulletin of Mathematical Biophysics*, 5, 115–133.

McFadden, D. (1973): "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka, Cambridge, Massachusetts: Academic Press, 105–142.

McGregor, C. (2015): "In-App Purchases of the Top Grossing Apps," *Growthbug Blog*, `https://growthbug.com/in-app-purchases-of-the-top-grossing-apps-db091584b69` (Accessed May 18 2020).

Mela, C. F., S. Gupta, and D. R. Lehmann (1997): "The Long-Term Impact of Promotion and Advertising on Consumer Brand Choice," *Journal of Marketing Research*, 248–261.

Milgrom, P. and J. Roberts (1986): "Price and Advertising Signals of Product Quality," *Journal of Political Economy*, 94, 796–821.

Milošević, M., N. Živić, and I. Andjelković (2017): "Early Churn Prediction with Personalized Targeting in Mobile Social Games," *Expert Systems with Applications*, 83, 326–332.

Misra, K., E. M. Schwartz, and J. Abernethy (2019): "Dynamic Online Pricing with Incomplete Information Using Multi-Armed Bandit Experiments," *Marketing Science*, 38, 226–252.

Moe, W. W. (2003): "Buying, Searching, or Browsing: Differentiating between Online Shoppers Using In-Store Navigational Clickstream," *Journal of Consumer Psychology*, 13, 29–39.

Moe, W. W. and P. S. Fader (2004): "Dynamic Conversion Behavior at E-Commerce Sites," *Management Science*, 50, 326–335.

Moro, S., P. Cortez, and P. Rita (2015): "Using Customer Lifetime Value and Neural Networks to Improve the Prediction of Bank Deposit Subscription in Telemarketing Campaigns," *Neural Computing and Applications*, 26, 131–139.

Müller-Navarra, M., S. Lessmann, and S. Voss (2015): "Sales Forecasting with Partial Recurrent Neural Networks: Empirical Insights and Benchmarking Results," in *2015 48th Hawaii International Conference on System Sciences*, IEEE, 1108–1116.

Nair, H. (2007): "Intertemporal Price Discrimination with Forward-Looking Consumers: Application to the US Market for Console Video-Games," *Quantitative Marketing and Economics*, 5, 239–292.

Nair, H. S., S. Misra, W. J. Hornbuckle, R. Mishra, and A. Acharya (2017): "Big Data and Marketing Analytics in Gaming: Combining Empirical Models and Field Experimentation," *Marketing Science*, 36, 699–725.

Neslin, S. A. and H. J. van Heerde (2009): "Promotion Dynamics," *Foundations and Trends® in Marketing*, 3, 177–268.

Netzer, O., J. M. Lattin, and V. Srinivasan (2008): "A Hidden Markov Model of Customer Relationship Dynamics," *Marketing Science*, 27, 185–204.

Nevskaya, Y. and P. Albuquerque (2019): "How Should Firms Manage Excessive Product Use? A Continuous-Time Demand Model to Test Reward Schedules, Notifications, and Time Limits," *Journal of Marketing Research*, 56, 379–400.

Niculescu, M. F. and D. J. Wu (2014): "Economics of Free under Perpetual Licensing: Implications for the Software Industry," *Information Systems Research*, 25, 173–199.

Nijs, V. R., M. G. Dekimpe, J.-B. E. Steenkamps, and D. M. Hanssens (2001): "The Category-Demand Effects of Price Promotions," *Marketing Science*, 20, 1–22.

Odlyzko, A. (2004): "Privacy, Economics, and Price Discrimination on the Internet," in *Economics of Information Security*, Berlin, Germany: Springer, 187–211.

Padilla, N. and E. Ascarza (2019): "The Value of First Impressions: Leveraging Acquisition Data for Customer Management," *Harvard Business School Working Paper 19-091*, `https://www.hbs.edu/faculty/Pages/item.aspx?num=55634` (Accessed May 18 2020).

Pauwels, K. and A. Weiss (2008): "Moving from Free to Fee: How Online Firms Market to Change Their Business Model Successfully," *Journal of Marketing*, 72, 14–31.

Perez, S. (2019): "Tinder Becomes the Top-Grossing, Non-Game App in Q1 2019, Ending Netflix's Reign," *Techcrunch*, `https://techcrunch.com/2019/04/11/tinder-becomes-the-top-grossing-non-game-app-in-q1-2019-ending-netflixs-reign` (Accessed May 18 2020).

Periáñez, Á., A. Saas, A. Guitart, and C. Magne (2016): "Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 564–573.

Petro, G. (2019): "Amazon's Crisis of Trust," *Forbes Online*, `https://www.forbes.com/sites/gregpetro/2019/03/08/amazons-crisis-of-trust/` (Accessed May 18 2020).

Pigou, A. (2017): *The Economics of Welfare*, Abingdon, United Kingdom: Routledge.

Pivetta, E., L. Harkin, J. Billieux, E. Kanjo, and D. J. Kuss (2019): "Problematic Smartphone Use: An Empirically Validated Model," *Computers in Human Behavior*.

Rana, R. and F. S. Oliveira (2015): "Dynamic Pricing Policies for Interdependent Perishable Products or Services Using Reinforcement Learning," *Expert Systems with Applications*, 42, 426–436.

Rao, A. R. and K. B. Monroe (1989): "The Effect of Price, Brand Name, and Store Name on Buyers' Perceptions of Product Quality: An Integrative Review," *Journal of Marketing Research*, 26, 351–357.

Reinartz, W. J. and V. Kumar (2003): "The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration," *Journal of Marketing*, 67, 77–99.

Ross, N. (2018): "Customer Retention in Freemium Applications," *Journal of Marketing Analytics*, 6, 127–137.

Rossi, P. E., R. E. McCulloch, and G. M. Allenby (1996): "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15, 321–340.

Rothenbuehler, P., J. Runge, F. Garcin, and B. Faltings (2015): "Hidden Markov Models for Churn Prediction," in *2015 SAI Intelligent Systems Conference (IntelliSys)*, IEEE, 723–730.

Rothschild, M. (1974): "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, 9, 185–202.

Rubin, D. B. (1978): "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.

Runge, J., P. Gao, F. Garcin, and B. Faltings (2014): "Churn Prediction for High-Value Players in Casual Social Games," in *Computational Intelligence and Games (CIG), 2014 IEEE Conference on*, IEEE, 1–8.

Runge, J., H. Nair, and J. Levav (2019): "Price Promotions in 'Freemium' Settings," *Stanford Graduate School of Business Working Paper No. 3769*, https://www.gsb.stanford.edu/faculty-research/working-papers/price-promotions-freemium-settings (Accessed May 18 2020).

Runge, J., S. Wagner, J. Claussen, and D. Klapper (2016): "Freemium Pricing: Evidence from a Large-Scale Field Experiment," *SSRN Working Paper 2888471*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2888471 (Accessed May 18 2020).

Sahni, N. S., D. Zou, and P. K. Chintagunta (2016): "Do Targeted Discount Offers Serve as Advertising? Evidence from 70 Field Experiments," *Management Science*, 63, 2688–2705.

Santosa, F. and W. W. Symes (1986): "Linear Inversion of Band-Limited Reflection Seismograms," *SIAM Journal on Scientific and Statistical Computing*, 7, 1307–1330.

Schmittlein, D. C., D. G. Morrison, and R. Colombo (1987): "Counting Your Customers: Who Are They and What Will They Do Next?" *Management Science*, 33, 1–24.

Schonfeld, E. (2009): "Pinch Media Data Shows the Average Shelf Life of an iPhone App Is Less Than 30 Days," *Techcrunch*, https://techcrunch.com/2009/02/19/pinch-media-data-shows-the-average-shelf-life-of-an-iphone-app-is-less-than-30-days/ (Accessed May 18 2020).

Schütze, A. (2014): "Zigaretten-Einzelverkauf Beliebt und Bald Verboten," *Stern magazine*, https://www.stern.de/wirtschaft/geld/tabakhandel-zigaretten-einzelverkauf-beliebt-und-bald-verboten-3070642.html (Accessed May 18 2020).

Schwartz, E. M., E. T. Bradlow, and P. S. Fader (2017): "Customer Acquisition via Display Advertising Using Multi-Armed Bandit Experiments," *Marketing Science*, 36, 500–522.

Sensortower (2019a): "Global App Revenue Grew 23% in 2018 to More Than 71 Billion USD on iOS and Google Play," *Sensortower*, `https://sensortower.com/blog/app-revenue-and-downloads-2018` (Accessed May 18 2020).

——— (2019b): "Q2 2019 Q2 2019 Store Intelligence Data Digest," *Sensortower*, `https://sensortower.com/reports` (Accessed May 18 2020).

Seufert, E. B. (2013): *Freemium Economics: Leveraging Analytics and User Segmentation to Drive Revenue*, Amsterdam, Netherlands: Elsevier.

Shampanier, K., N. Mazar, and D. Ariely (2007): "Zero as a Special Price: The True Value of Free Products," *Marketing Science*, 26, 742–757.

Shapiro, C. (1983): "Optimal Pricing of Experience Goods," *The Bell Journal of Economics*, 14, 497–507.

Shapiro, C. and H. R. Varian (1998): *Information Rules: A Strategic Guide to the Network Economy*, Brighton, Massachusetts: Harvard Business Press.

Shi, S. W., M. Xia, and Y. Huang (2015): "From Minnows to Whales: An Empirical Study of Purchase Behavior in Freemium Social Games," *International Journal of Electronic Commerce*, 20, 177–207.

Shi, Z., K. Zhang, and K. Srinivasan (2019): "Freemium as an Optimal Strategy for Market Dominant Firms," *Marketing Science*, 38, 150–169.

Shiller, B. R. (2020): "Approximating Purchase Propensities and Reservation Prices from Broad Consumer Tracking," *International Economic Review*, 61, 847–870.

Shmilovici, U. (2011): "The Complete Guide To Freemium Business Models," *Techcrunch*, `https://techcrunch.com/2011/09/04/complete-guide-freemium/` (Accessed May 18 2020).

Sifa, R., F. Hadiji, J. Runge, A. Drachen, K. Kersting, and C. Bauckhage (2015): "Predicting Purchase Decisions in Mobile Free-to-Play Games," *Proceedings of AAAI Articial Intelligence in Interactive Digital Entertainment Conference*.

Sifa, R., J. Runge, C. Bauckhage, and D. Klapper (2018): "Customer Lifetime Value Prediction in Non-Contractual Freemium Settings: Chasing High-Value Users Using Deep Neural Networks and SMOTE," in *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Sinclair, B. (2017): "Zynga Apologizes for Random DLC Pricing Experiment," *Gamesindustry Biz*, `http://www.gamesindustry.biz/articles/2017-07-14-zynga-apologizes-for-random-dlc-pricing-experiment` (Accessed May 18 2020).

Sivaramakrishnan, S., F. Wan, and Z. Tang (2007): "Giving an 'E-Human Touch' to E-Tailing: The Moderating Roles of Static Information Quantity and Consumption Motive in the Effectiveness of an Anthropomorphic Information Agent," *Journal of Interactive Marketing*, 21, 60–75.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014): "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, 15, 1929–1958.

Steinkuehler, C. A. (2004): "Learning in Massively Multiplayer Online Games," in *Proceedings of the 6th International Conference on Learning Sciences*, International Society of the Learning Sciences, 521–528.

Surrette, B. (2019): "Training Sets, Validation Sets, and Holdout Sets," *DataRobot Wiki*, `https://www.datarobot.com/wiki/training-validation-holdout/` (Accessed May 18 2020).

Sutton, R. S. and A. G. Barto (1999): "Reinforcement Learning," *Journal of Cognitive Neuroscience*, 11, 126–134.

——— (2018): *Reinforcement Learning: An Introduction*, Cambridge, Massachusetts: MIT Press.

The World Bank (2018): "World Development Indicators," *World Bank Data Catalog*, `https://datacatalog.worldbank.org/dataset/world-development-indicators` (Accessed April 14 2018).

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, 58, 267–288.

Tong, S., X. Luo, and B. Xu (2020): "Personalized Mobile Marketing Strategies," *Journal of the Academy of Marketing Science*, 48, 64–78.

Trusov, M., R. E. Bucklin, and K. Pauwels (2009): "Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site," *Journal of Marketing*, 73, 90–102.

Turban, E. and P. R. Watkins (1986): "Integrating Expert Systems and Decision Support Systems," *Management Information Systems Quarterly (MISQ)*, 10, 121–136.

Vanderveld, A., A. Pandey, A. Han, and R. Parekh (2016): "An Engagement-Based Customer Lifetime Value System for E-Commerce," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 293–302.

Varian, H. R. (1980): "A Model of Sales," *American Economic Review*, 70, 651–659.

Verhoef, P. C., P. N. Spring, J. C. Hoekstra, and P. S. Leeflang (2003): "The Commercial Use of Segmentation and Predictive Modeling Techniques for Database Marketing in the Netherlands," *Decision Support Systems*, 34, 471–481.

Vincent, P., H. Larochelle, Y. Bengio, and P.-A. Manzagol (2008): "Extracting and Composing Robust Features with Denoising Autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, ACM, 1096–1103.

Voigt, S. and O. Hinz (2016): "Making Digital Freemium Business Models a Success: Predicting Customers' Lifetime Value via Initial Purchase Information," *Business & Information Systems Engineering*, 58, 107–118.

Von Neumann, J. and O. Morgenstern (2007): *Theory of Games and Economic Behavior (Commemorative Edition)*, Princeton, New Jersey: Princeton University Press.

Wager, S. and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113.

Wang, Y., M. Lewis, and V. Singh (2016): "The Unintended Consequences of Countermarketing Strategies: How Particular Anti-Smoking Measures May Shift Consumers to More Dangerous Cigarettes," *Marketing Science*, 35, 55–72.

Weiss, G. M. (2004): "Mining with rarity: a unifying framework," *ACM SIGKDD Explorations Newsletter*, 6, 7–19.

West, P. M., P. L. Brockett, and L. L. Golden (1997): "A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice," *Marketing Science*, 16, 370–391.

Xie, Y., X. Li, E. Ngai, and W. Ying (2009): "Customer Churn Prediction Using Improved Balanced Random Forests," *Expert Systems with Applications*, 36, 5445–5449.

Yadav, S. and S. Shukla (2016): "Analysis of K-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, IEEE, 78–83.

Yeykelis, L., J. J. Cummings, and B. Reeves (2014): "Multitasking on a Single Device: Arousal and the Frequency, Anticipation, and Prediction of Switching Between Media Content on a Computer," *Journal of Communication*, 64, 167–192.

——— (2018): "The Fragmentation of Work, Entertainment, E-Mail, and News on a Personal Computer: Motivational Predictors of Switching Between Media Content," *Media Psychology*, 21, 377–402.

Zenetti, G., T. H. Bijmolt, P. S. Leeflang, and D. Klapper (2014): "Search Engine Advertising Effectiveness in a Multimedia Campaign," *International Journal of Electronic Commerce*, 18, 7–38.

Zhang, G., B. E. Patuwo, and M. Y. Hu (1998): "Forecasting with Artificial Neural Networks: The State of the Art," *International Journal of Forecasting*, 14, 35–62.

Zhang, J. and M. Wedel (2009): "The Effectiveness of Customized Promotions in Online and Offline Stores," *Journal of Marketing Research*, 46, 190–206.

Zucchini, W., I. L. MacDonald, and R. Langrock (2017): *Hidden Markov Models for Time Series: An Introduction Using R*, Boca Raton, Florida: CRC press.

Selbstständigkeitserklärung

Ich versichere, die von mir vorgelegte Dissertation selbständig und ohne unerlaubte Hilfe und Hilfsmittel angefertigt sowie die benutzten Quellen und Daten anderen Ursprungs als solche kenntlich gemacht zu haben.
Ich bezeuge durch meine Unterschrift, dass meine Angaben über die bei der Abfassung meiner Dissertation benutzten Hilfsmittel, über die mir zuteil gewordene Hilfe sowie über frühere Begutachtungen meiner Dissertation in jeder Hinsicht der Wahrheit entsprechen.

San Francisco, den 25. Mai 2020

Julian Runge